Processus de rang et applications statistiques en grande dimension

Myrto Limnios

université
PARIS-SACLAY

Processus de Rang et Applications Statistiques en Grande Dimension

Rank Processes and Statistical Applications in High Dimension

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°574 : Mathématiques Hadamard (EDMH)
Spécialité de doctorat: Mathématiques aux interfaces
Graduate School : Mathématiques, Référent : ENS Paris-Saclay

Thèse préparée dans l'unité de recherche Centre Borelli (Université Paris-Saclay, CNRS, ENS Paris-Saclay), sous la direction de Nicolas VAYATIS, le co-encadrement de Ioannis BARGIOTAS

Thèse soutenue à Paris-Saclay, le 14 mars 2022, par

Myrto LIMNIOS

Composition du jury

Pascal Massart Professeur, Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay, France	Président
Alexandra Carpentier Professeure, Institut de mathématiques, Université de Potsdam, Allemagne	Rapporteur & Examinatrice
Johan Segers Professeur, Institut de statistique, biostatistique et sciences actuarialles, LIDAM, UCLouvain, Belgique	Rapporteur & Examinateur
Stephan Clémençon Professeur, Telecom Paris, LTCI, Institut Polytechnique de Paris, France	Examinateur
Sara van de Geer Professeure, Séminaire de statistiques, Département de mathématiques, ETH Zürich, Suisse	Examinatrice
Nicolas Vayatis Professeur, Université Paris-Saclay, ENS Paris-Saclay, CNRS, Centre Borelli, France	Directeur de thèse

THESE DE DOCTORAT
NNT : 2022UPASM006

Contents

Remerciements

Mes premiers remerciements s'adressent à mon directeur de thèse Nicolas Vayatis. Sa confiance et sa générosité m'ont permise de commencer ce projet de thèse et de me plonger dans un riche environnement de recherche. Il m'a toujours encouragée à participer aux différents séminaires et conférences scientifiques en lien avec mes travaux. Son exigence n'a cessé de pousser mes limites et de rafiner ma critique scientifique. J'espère garder de nos discussions une ouverture d'esprit et une rigueur liées à l'utilisation de méthodes statistiques et d'apprentissage. Cette thèse n'aurait pu avoir lieu sans son soutient. Merci à Ioannis Bargiotas pour son encadrement durant mes débuts. Il a su m'intégrer aux projets biomédicaux, me sensibiliser aux problématiques et méthodologies liées à l'étude de données humaines.

Je voudrais particulièrement remercier Stephan Clémençon, qui a été un pilier tout au long de cette thèse. Il m'a appris à aller à l'essentiel tout en ne cédant pas à la facilité. Je le remercie sincèrement pour sa patience et sa pédagogie, qui m'ont initiée à ce métier de chercheuse. Je souligne sa volonté de m'intégrer dans son équipe de recherche me permettant ainsi d'échanger avec les autres doctorants et chercheurs. Enfin, pour la confiance qu'il m'a accordée dans mes missions d'enseignement.

Je remercie les membres de mon jury pour avoir accepté cette tâche, et en premier temps, mes deux rapporteurs : Alexandra Carpentier et Johan Segers. Je suis honorée de leurs investissement et curiosité, je n'aurais pu espérer des retours plus complets et lumineux sur ce travail. Je remercie sincèrement Pascal Massart pour avoir accepté de présider le jury, et à Sara van de Geer pour avoir été membre. Jamais je n'aurais pensé avoir la chance et l'honneur de présenter mon travail à ce jury. Enfin, il est à noter également que leurs travaux ont été fondamentaux pour les recherches présentées dans ce manuscrit.

Cette thèse fut aussi un long voyage qui m'a offert de nombreuses opportunités de rencontres. Je tiens donc à remercier celles et ceux avec qui j'ai eu la chance de partager ce quotidien, les joies mais aussi les moments plus laborieux, notamment dans ce contexte de crise sanitaire (Alice et Marie !). J'espère vous recroiser de nouveau. A mes merveilleux co-bureaux : Etienne, Flore et Thomas; au groupe du CMLA : Alejandro, Alice, Anthéa, Antoine, Ayman, Batiste, Brian, Cédric, Charles, Firas, Ludo, Marie, Mathilde, Matthieu, Mounir, Pierre, Quentin, Tina, Tristan; et aux nouveaux ! Mais aussi ceux qui m'ont accueillie à Télécom : Alex, Guillaume, Kevin, Mastane, Mathurin, Nathan, Nidham, Pierre, Pierre A. et L., Robin, Yannick... Et aux autres : Anna, Gauthier, ... Une mention toute particulière à ceux de la dernière ligne droite (Alice, Firas et Nidham !) au souvenir de nos journées de rédaction passées à la BNF, heureusement que vous étiez là !

Merci aussi à Vianney Perchet, qui m'a intégrée dans son groupe de lecture, alors que nos sujets de recherche n'étaient a priori pas communs. Merci à toute l'équipe administrative et logistique du Centre Borelli pour leurs bonne humeur et aide essentielle, et en particulier à Alina Muller, Christophe Labourdette, Gwladys Stouvenel, Véronique Almadovar et Virginie Pauchont.

Je voudrais finir par remercier mes proches et amis pour leurs soutient et optimisme depuis le

début, quelle chance de grandir à vos côtés ! Nos rendez-vous, malgré la distance pour certains, sont indispensables et j'espère qu'ils continueront. Merci aussi à mes professeurs de danse, pour l'équilibre essentiel apporté au travers de cette pratique. Merci à Alexandre, dont le soutient et la patience ont été précieux et, pour qui j'espère, l'aboutissement de ce travail en donnera un sens. Enfin, mes derniers mots s'adressent à mes parents et mon frère, pour leur soutient exceptionnel et inébranlable. Vous m'avez tant appris, avec bienveillance et générosité rares.

Notation

General

d	Dimension of an Euclidean feature space, valued in $\mathbb{N}^*$	
X, Y, Z	Univariate random variables	p. 9
$\mathbf{X}, \mathbf{Y}, \mathbf{Z}$	Multivariate random variables	p. 4
$\mathbf{1}_n$	Unit n-dimensional vector, $n \in \mathbb{N}^*$	p. 131
$\mathbb{I}_d$	Identity square matrix of dimension $d \times d$	p. 119
$S_d^+(\mathbb{R})$	Set of positive definite matrices of dimension $d \times d$	p. 119
$\mathbb{I}\{\mathscr{E}\}$	Indicator function for an event $\mathscr{E}$	p. 9
δ_x	Dirac mass at any point x	p. 4
$\mathcal{S}_n$	Set of all permutations on $\{1, \ldots, n\}$	p. 57
$\mathscr{Z}$	Measurable multi-dimensional (feature) space	p. 1

Distributions and functions

$\lfloor \cdot \rfloor$	Floor function	p. 9
$\lceil \cdot \rceil$	Ceiling function	p. 9
$\|f\|_\infty = \sup_{x \in \mathscr{Z}} \|f(x)\|$	Sup norm for a function $f : \mathscr{Z} \to \mathbb{R}$	
$\|\cdot\|$	Euclidean norm in $\mathbb{R}^d$, $d \in \mathbb{N}^*$	p. 33
$\mathscr{F}$	Class of measurable functions $f : \mathscr{Z} \to \mathbb{R}$	p. 54
$\|\mu\|_{\mathscr{F}} = \sup_{f \in \mathscr{F}} \|\mu(f)\|$	Sup norm of $\mathscr{F}$ *w.r.t.* a measure μ	p. 54
$\mathcal{S}^{d-1} = \{z \in \mathscr{Z},\ \|z\| = 1\}$	Unit sphere on $\mathscr{Z} \subset \mathbb{R}^d$	p. 32
$W^{-1}(u) = \inf\{t \in]-\infty,\ +\infty] :\ W(t) \geq u\}$	Generalized inverse/pseudo-inverse of a cumulative distribution càd-làg function $W(\cdot)$ on $\mathbb{R} \cup \{+\infty\}$	p. 38
$\lambda(\cdot)$	Lebesgue measure	p. 33
$\mathscr{U}_d(\cdot)$	Uniform distribution on a set $\mathscr{Z} \subset \mathbb{R}^d$	p. 154
$\mathscr{B}(\cdot)$	Bernoulli distribution	p. 45
$\mathscr{N}_d(\cdot, \cdot)$	Gaussian distribution of dimension d	p. 119

Rank statistics

p	Asymptotic proportion of a sample among the two samples, in $(0,1)$	p. 9,37
$\phi(\cdot)$	Score-generating function $[0,1] \to \mathbb{R}$	p. 9
$\mathrm{Rank}(\cdot)$	Simple rank statistic	p. 9
$s(\cdot)$	Scoring function, measurable $\mathscr{Z} \to]-\infty, \infty]$	p. 13
$\mathcal{S}$	Class of scoring functions s	p. 13
$\mathscr{V}$	Vapnik-Chervonenkis dimension of a class	p. 55
$\Psi(\cdot)$	Likelihood ratio function	p. 39
$\eta(\cdot)$	Posterior probability	p. 37

Hypothesis testing

$\mathscr{H}_0$	Null hypothesis	p. 4
$\mathscr{H}_1$	Alternative hypothesis	p. 4
α	Level/size of a statistical test in $(0,1)$	p. 4

Acronyms

ROC	Receiver Operating Characteristic (curve)	p. 38
AUC	Area Under the ROC Curve	p. 40
MV	Mass-Volume (curve)	p. 47
ERM	Empirical Risk Minimization	p. 50
VC	Vapnik-Chervonenkis	p. 55
SVM	Support Vector Machine	p. 13
NN	Neural Net	p. 13
MWW	Mann-Whitney-Wilcoxon	p. 77
RTB	Ranking-the-best	p. 78
MMD	Maximum Mean Discrepancy	p. 34
CoP	Center of Pressure	p. 184
CoM	Center of Mass	p. 185

“Lock up your libraries if you like; but there is no gate, no lock, no bolt that you can set upon the freedom of my mind.”

V. Woolf, A Room of One’s Own.

1 | Introduction

"All knowledge-beyond that of bare isolated occurrence-deals with uniformities. Of the latter, [...] the vast majority are only partial; medicine does not teach that smallpox is inevitably escaped by vaccination, but that it is so generally; biology has not shown that all animals require organic food, but that nearly all do so; in daily life, a dark sky is no proof that it will rain, but merely a warning; even in morality, the sole categorical imperative alleged by Kant was the sinfulness of telling a lie, and few thinkers since have admitted so much as this to be valid universally."

C. Spearman, The Proof and Measurement of Association Between Two Things.

1.1 Context and motivations

The high-dimensional and nonparametric two-sample problem. In its most general statistical formulation, the two-sample problem tests the equality of two unknown probability distributions at a level of risk, when considering two independent *i.i.d.* random samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, valued on the (same) measurable space $\mathscr{Z}$, for instance of $\mathbb{R}^d$, $d \geq 2$. While there is long-standing literature for the univariate setting (see Lehmann and Romano (2005)), this problem remains a research subject for both the multivariate and nonparametric frameworks. Indeed, the increasing ability to collect large, even massive data, of various structure that is possibly biased due to the collection process, has strongly defied classic modelings, see *e.g.* Wang et al. (2019). Such types of data are in particular analyzed in applied fields as in biomedicine (*e.g.* clinical trials, genomics), in marketing (*e.g.* A/B testing, recommendation systems), in economics, *etc.* The recent methods for high-dimensional setting usually rely on distance-based statistics. These distances are estimated on empirical versions of the underlying probability measures (or related), such that the more the distance decreases, the more the two samples can be characterized as homogeneous, see Biau and Gyorfi (2005); Gretton et al. (2012a). Unfortunately, these formulations often depend on the intrinsic characterization of the metric, and on the ambient representation of the random observations. Additionally, they often lack certain important statistical properties regarding, for instance, the nonasymptotic control of the type-I and/or type-II errors, the computation of the exact null distribution, or even the stability *w.r.t.* the dimension of the space $\mathscr{Z}$ (*e.g.* of d). In the high-dimensional case, most work have considered semiparametric statistical models and focus on the location or a scale tests, see *e.g.* Baringhaus and Franz (2004).

Rank statistics and learning-to-rank methods. Going back to Spearman's rho test published in Spearman (1904), rank statistics were introduced as a response to the traditional 'Gaussian assumption'. Precisely, as the observations are solely considered through their relative order, *R*-statistics *"reduce the "accidental errors""* (page 81, Spearman (1904)). Later, they gained popularity thanks to their simplicity, fast computation, and being a particular class of permutation statistics. In the context of two-sample testing, it is among the most competitive class of test statistics under mild conditions on the underlying probability distributions, see *e.g.* Chernoff and Savage (1958); Hodges and Lehmann (1956). They are proved to achieve exact distribution-free distribution under the null hypothesis (equality of distributions) while providing high power in the univariate framework, see Chap. 15 in van der Vaart (1998). The simplest version, known as the *ranksum* or Mann-Whitney-Wilcoxon test statistic (Mann and Whitney (1947); Wilcoxon (1945)), is celebrated for being asymptotically uniformly most powerful for the location problem at fixed test size and under logistic distributions, see Ex.15.15 in van der Vaart (1998). However, the definition of rank statistics is far from being straightforward for high-dimensional settings due to the lack of natural order in multivariate data. The literature usually relies on depth or spatial ranks that heavily depend on their intrinsic definition and are mainly designed for particular (parametric) tests, see *e.g.* Chakraborty and Chaudhuri (2017); Hallin and Paindaveine (2008).

In a different context, ranking observations has become fundamental in many data analysis problems for the past decades, *e.g.*, in information retrieval and computational biology. They are defined as *learning-to-rank* methods, and aim to learn an order from a set of observations according to their relevance/importance/preference to predict the order of any 'new' data sample. In particular, the most simple approach for two samples is known to be intimately related to the Mann-Whitney-Wilcoxon statistic *via* the *Receiver Operating Characteristic* (ROC) analysis, see Clémençon and Vayatis (2009b). A series of fundamental contributions relating learning-to-rank approaches and ROC analysis are, for instance, Agarwal et al. (2005); Cortes and Mohri (2004). However, and to the best of our knowledge, only Clémençon et al. (2008, 2009) theoretically leveraged the relation to classes of some linear rank statistics. The latter direction motivates this work and ideally seeks algorithmic procedures that are easy-to-use, interpretable, and trackable.

Towards nonasymptotic guarantees. More generally, complex data structures with possibly biased acquisition conditions require nonparametric and multivariate statistical modeling. In hypothesis testing, standard multivariate extensions of classic univariate test statistics lack nonparametric analysis. Valiant (1984) introduced the *Probabilistic Approximately Correct* (PAC) theory, providing a framework for quantifying the difficulty of a data analysis problem. Briefly, by considering a probabilistic space $(\Omega, \mathscr{A}, \mathbb{P})$, PAC bounds formally aim to control, at a certain probability, an event A as follows:

$$\textit{For fixed probability } \delta > 0, \textit{ for any element } \omega \in A(\delta) \textit{ such that the event } A(\delta) \subset \Omega, \\ \textit{satisfies } \mathbb{P}\{A(\delta)\} \geq 1 - \delta.$$

Actually, tail bounds could be derived for an estimator Z_N based on a sample of size $N \in \mathbb{N}^*$, as finding the threshold $t_{\delta,N} > 0$ such that $\mathbb{P}\{Z_N < t_{\delta,N}\} \geq 1 - \delta$, where $A(\delta) = \{Z_N < t_{\delta,N}\}$. With probability $1 - \delta$, it ensures that the random variable does not exceed a certain threshold, incidentally interpreted as a (nonasymptotic) confidence interval of the estimator. This type of bounds is particularly used in statistical learning theory for studying the random fluctuations of the empirical risk given a model.

Biomedical application: comparison of posturographic data. This thesis is motivated by a biomedical project regarding the quantification of human behavior developed in an interdisciplinary

research team[1]. In particular, in the context of personalized medicine and elderly prevention, a team gathering mathematicians, (bio)statisticians, and clinicians from different specialties studies the postural control of clinical populations. The idea is to be able to collect the evolution of the postural control through clinical follow-ups to detect a possible deterioration. Specifically, frailty progression in late adulthood for Parkinsonian patients is at the heart of the project. This population is more subject to postural instability, implying possible falls at ages for which surgical operations are not encouraged. A way of measuring the postural control is by using sensorimotor platforms that register during a short timescale the temporal variation of the *Center of Pressure* (CoP) displacement (statokinesigram) of the patient. The experimental protocol is illustrated in Figure 1.1.

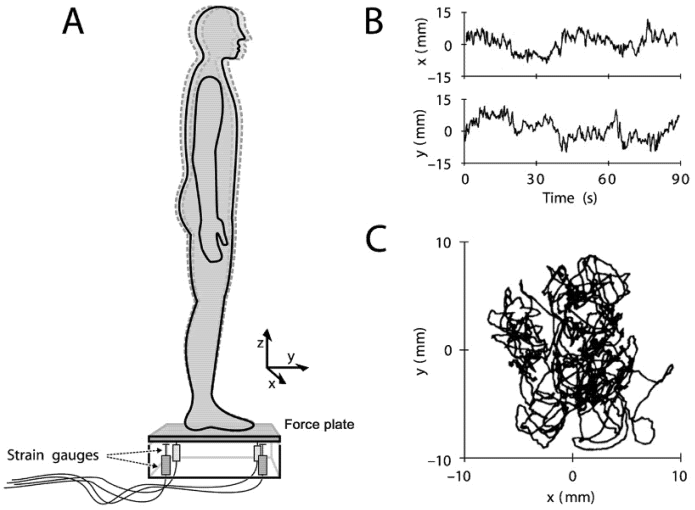


Figure 1.1. Illustration of the statokinesigram acquisition protocol. The patient stands still in (A) to measure the two-dimensional trajectory of the center of pressure by the force platform. The two timeseries of the Medio-Lateral and the Antero-Posterior are *resp.* defined as the x-axis and the y-axis in (B). An example of statokinesigram in (C). Source: Chen et al. (2021).

In this context, a typical and important problem renders in the comparison of patients having frail postural control, referred to as *Faller*, *w.r.t.* a chosen 'control' population, referred to as *Non Faller*. To better understand the difficulty of this question, Figure 1.2 gathers statokinesigrams measured from these two populations, for which the visual distinction between the pairs of patients (a *vs.* b, and c *vs.* d) is far from being straightforward. The measurements are of complex structure (*e.g.* multiple features, functional nature, small/imbalanced cohorts), for which additional information/features about the patients can be added to (*e.g.* comorbidity, age). In fact, after adequate preprocessing, many characteristics from the obtained statokinesigrams can be collected for the analysis, see Quijoux et al. (2021). However, there are strong limitations when using traditional two-sample testing approaches to such data types. Practitioners face either approach challenging to implement or univariate and parametric models that are not adequate. We explored typical ones in Bargiotas et al. (2021) in the context of postural control, and more generally highlighted some scientific facts raised by the scientific community for the use of statistics in Appendix Chap. C.1. For instance, multiple testing procedures are usually associated with simple corrections controlling the type-I error for comparing multivariate observations, see *e.g.* Hochberg (1988); Hommel (1988). We compared their ability to discriminate those two populations (*Faller*/*Non Faller*) to two multivariate methods: the

[1]The Centre Borelli is a research laboratory resulting from the recent merger of an applied mathematics laboratory (CMLA, Ecole Normale Supérieure Paris-Saclay) and a neuroscience laboratory (COGNAC-G, Université Paris Descartes), wherein multiple interdisciplinary projects are developed.

Maximum Mean Discrepancy (MMD), see Gretton et al. (2007), and the proposed generalization of rank statistics based on Clémençon et al. (2009). Classical procedures failed to reject the null hypothesis, *i.e.* , concluded that both populations are drawn from the same distribution. On the contrary, both multivariate methods concluded in a significant difference with very low p-values. We refer to Chapter 9 for the detailed results, Tables 9.3 and 9.4 therein.

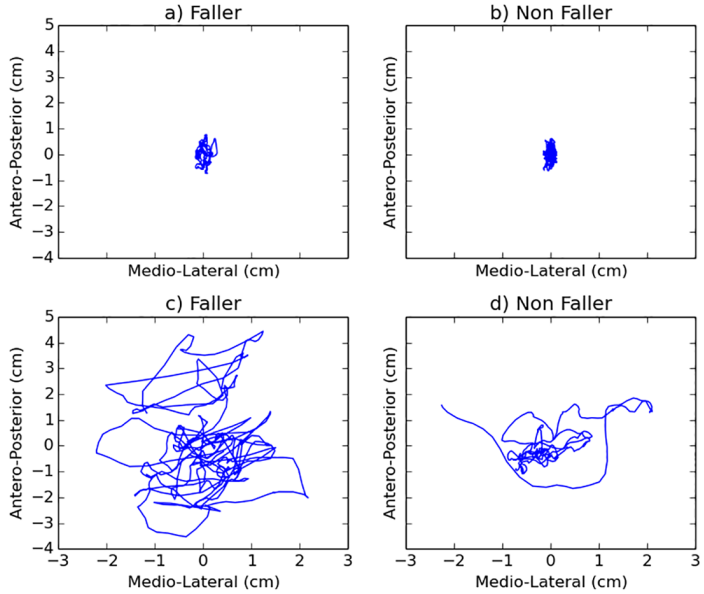


Figure 1.2. Illustration of statokinesigrams for *Faller* (a,c) and *Non-Faller* (b,d) patients in the Medio-Lateral/Antero-Posterior space. Source: Audiffren et al. (2016).

1.2 The high-dimensional two-sample problem

This section formulates the two-sample problem in the multivariate and nonparametric setting. In particular, state-of-the-art statistical tests are reviewed, while some limitations are subsequently discussed. We refer to the Appendix section B.2 for its univariate formulation with a review on classical properties and statistics.

1.2.1 Formulation

Consider two independent random variables $\mathbf{X}$ and $\mathbf{Y}$, defined on a probability space and valued in the (same) multivariate measurable space $\mathscr{Z}$, of unknown continuous distribution functions G and H. For a fixed level $\alpha \in (0,1)$, the two-sample problem corresponds to testing the two hypothesis below:

$$\mathscr{H}_0 : \; G = H \text{ against the alternative } \mathscr{H}_1 : \; G \neq H \; . \tag{1.2.1}$$

Also known as homogeneity testing, many classic statistical problems can be related to. See Darling (1957) for the univariate goodness-of-fit testing and Friedman (2004) for the multivariate model, Spearman (1904) for independence testing, and Wilcoxon (1945) for pairwise testing.

In practice, and especially for nonparametric settings, we consider independent copies of the *r.v.* as the underlying (classes of) distributions are unknown. Let $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$, with $n, m \in \mathbb{N}^*$, two independent *i.i.d.* samples drawn from G and H, and valued in the (same) measurable space $\mathscr{Z}$. Univariate nonparametric statistics, *e.g.*, Kolmogorov-Smirnov statistic (Smirnov (1939)), rely on empirical estimates of the underlying distributions or related (pseudo)-metrics, see Appendix

section B.2. The null hypothesis $\mathscr{H}_0$ is rejected if obtaining 'large' values of these statistics, *i.e.*, under 'large deviations' of the two random samples. For multivariate observations, natural empirical counterparts of their distributions are, for instance,

$$\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbf{X}_i} \text{ and } \widehat{\nu}_m = \frac{1}{m}\sum_{j=1}^{m} \delta_{\mathbf{Y}_j} , \tag{1.2.2}$$

where δ_x is the Dirac mass at any point x, or empirical versions of the *c.d.f.* , quantiles, copulas, depths, *etc.* Classic (pseudo-)metrics measuring dissimilarity between two probability distributions are: chi-square distance, Kullback-Leibler divergence, Hellinger distance, Kolmogorov-Smirnov distance. Refer to Rachev (1991) for a comprehensive review. In minimax testing, the alternative corresponds to the underlying distributions being different and separated in a metric sense, see *e.g.*, Lam-Weil et al. (2022) for local minimax separation rate defined by L_1-norm for discrete distributions, Carpentier et al. (2018) for L_2-norm in sparse linear regression. We refer in particular to Albert et al. (2021); Berrett et al. (2021) for independence testing, and to Baraud (2002); Ingster and Suslina (2003, 2000); Lepski and Spokoiny (1999) for goodness-of-fit testing. Lastly, a related problem in computer science literature refers to the two-sample problem as property testing, see for instance Goldreich et al. (1998); Rubinfeld and Sudan (1996). The example below formulates a classic statistical test known as the location test.

Example 1. (LOCATION TEST IN $\mathbb{R}^d$) *In (semi)parametric testing, by considering P_1, $P_2 \in \mathscr{P}$ a probabilistic model, such that $G(t) = P_1(t-\theta_1)$, $H(t) = P_2(t-\theta_2)$, with parameters θ_1, $\theta_2 \in \mathbb{R}^d$, with $d \in \mathbb{N}^*$, the location problem is formulated as*

$$\mathscr{H}_0 : \ \theta_1 = \theta_2 \quad vs. \quad \mathscr{H}_1 : \ \theta_1 \neq \theta_2 \ .$$

The simplest form is usually presented when supposing P_1, P_2 known and equal. It recovers the Hotelling's T^2-test for the equality of means for Gaussian distributions.

While statistics can be constructed for a particular probabilistic model, *e.g.* Gaussian, Elliptical models, this manuscript focuses on nonparametric formulations for which obtaining statistical guarantees is possible. Precisely, we are interested in (asymptotic) consistency, (asymptotic) control of both statistical errors (type-I and type-II), independence of the test statistics null distribution to the underlying model, independence of the test statistics to the transformations of the model under the alternative (also known as ancillary statistics Fisher (1925)), unbiasedness of the test statistic. Refer to Appendix section B.2 for details and definitions. Refer to classic books Gibbons and Chakraborti (2011); Lehmann and Romano (2005); Sheskin (2011); van der Vaart (1998) for comprehensive reviews of theory, methodologies and statistics in the field of (nonparametric) hypothesis testing.

1.2.2 State-of-the-art

This section reviews methods developed for the two-sample problem and formulated under nonparametric multivariate assumptions. The large majority is based on estimating a distance between the underlying probability measures (or related) of the two samples. The heuristic supposes the null hypothesis $\mathscr{H}_0$ to be equivalent to obtaining a zero distance. First, we present some extensions of classic univariate statistics, then detail new approaches, and finally, set out references for semiparametric models or applied to particular data structures. We refer to Section 2.1, in Chap. 2, for extensive reviews on the following methods.

Multivariate generalizations of classic univariate statistics. Bickel (1969) proposed a straightforward generalization of the Smirnov statistic (see Appendix B.2) and proved its distribution-free under the null. It seems as the first result to tackle such generic formulation without any additional sampling or post-analysis techniques. Then, Friedman and Rafsky (1979) obtained a graph-based generalization of both Wald-Wolfovitz runs statistic and Smirnov statistic (see Eq. (B.2.5), Wald and Wolfowitz (1940)). They constructed both statistics thanks to the subtree minimizing the total sum of interpoints distances, defined as the *Minimal Spanning Trees* (MST). The obtained statistics highly depend on both the chosen referential and distance, while being independent on the underlying distributions only when conditioned on the pooled samples.

Later, a series of extended generalizations proved advanced theoretical results using probability theory. Empirical processes defined collections of two-sample statistics, indexed by infinite-dimensional classes of functions of controllable complexity, for applying, *e.g.*, Glivenko-Cantelli and Donsker theorems. Three subsequent works generalized the Kolmogorov-Smirnov statistic, see Eq. (B.2.4) (Smirnov (1939)). First, Præstgaard (1995) proposed to map multivariate observations to the real line thanks to a scoring function, ranging in classes dependent on the sample's sizes. In Biau and Gyorfi (2005), the deviation between the two empirical measures is estimated on finite partitions of the ambient space. Recently, Zhou et al. (2017) introduced an alternative approach if the marginals are supposed to be independent. It is obtained by linear projections of the observations on functional decomposition basis. Additionally, Szekely and Rizzo (2004) and Baringhaus and Franz (2004) studied two versions for the multivariate energy statistic. While the first is straightforward, by considering the multivariate Euclidean distance, the second relies on linear projections on the unit sphere. Lastly, Clémençon et al. (2009) introduced a generalization of Mann-Whitney-Wilcoxon (MWW) statistic (see Eq. (B.2.3), Wilcoxon (1945)) using a bipartite ranking approach. This method learns the optimal mapping of the observations to the real line thanks to a scoring function, to induce a relation order on the feature space. The authors proved the asymptotic consistency of the procedure and presented promising numerical experiments.

Statistics based on kernel methods. This approach estimates dissimilarity measures in Hilbert space embeddings, by mapping a probability distribution into a *Reproducing Kernel Hilbert Space* (RKHS). The statistics are based on the mapped observations using kernel functions. Formalized in Gretton et al. (2007) and later in Gretton et al. (2012a), they proposed the *Maximum Mean Discrepancy* (MMD) statistic measuring the uniform bound in expectations over functions in the unit ball of a RKHS. In Gretton et al. (2012a), Theorem 5, they proved that if the RKHS is universal and the unit ball of class of functions is valued on a compact set, then the statistic equals zero *iff* the underlying distributions are equal. Considered as a classic approach, many developments have been published since, *e.g.*, Chwialkowski et al. (2016); Gretton et al. (2009, 2012b); Li et al. (2017); Schrab et al. (2021). While MMD is a metric in the uniform sense, Bach et al. (2008) considered a L_2-distance, which constructs a statistic generalizing Hotelling's T^2 test of the estimators based on kernel methods.

Statistics based on optimal transport distances. Methods relying on optimal transport theory compare probability measures in metric spaces *via* transport measures, such as the family of Wasserstein distances, see Villani (2009). Notice that the p-Wasserstein distance at power $p \in [1, \infty)$ is also known as the Mallow's distance in the statistical literature. In this line, Ramdas et al. (2015) defined the test statistic as the estimator minimizing an objective function: by writing the p-Wasserstein distance as the scalar product of the statistic and the pairwise distances between the two samples to the power p, penalized by the empirical entropy of the estimator. Recently, an approach at the crossroads of metric learning and rank statistics has been proposed, yielding to *population metric rank maps*,

see in particular Deb and Sen (2019). In Deb et al. (2021), authors extended classic statistics such as the energy or the T^2-Hotelling ones.

Further readings in semiparametric testing. Extensive and rich literature exists on semiparametric models, in particular for the location (Ex. 1) and scale (equality of variances testing) models under various assumptions *w.r.t.*, for instance, the structure of $\mathscr{Z}$ or the family of distributions (Gaussian, Elliptic, *etc.*). For completeness, we list additional concepts for the two-sample problem: nearest-neighbors tests Henze (1988); Schilling (1986), matching/assignment Mukherjee et al. (2020), permutation tests Hall and Tajvidi (2002), classifier Lopez-Paz and Oquab (2016), random projections Lopes et al. (2011); Srivastava et al. (2016), random forest Hediger et al. (2021), sparse mixture model Arias-Castro and Wang (2017), differential privacy Couch et al. (2019); Lam-Weil et al. (2020); Si et al. (2021). Additionally, Bhattacharya (2019); Lovato et al. (2020) review tests applied to graph structures or formulated *via* multivariate data-depths. See also Ingster and Suslina (2003) for a comprehensive overview of Gaussian models.

1.2.3 Limitations

This section focuses on (non)asymptotic statistical properties and possible associated algorithms to discuss on the approaches. Refer to Section 2.1, Chap. 2, for greater details.

We first review the results related to the computation or estimation of the null distribution, *i.e.*, when the underlying distributions G and H are supposed to be equal. Guarantees on the null distribution lead to a better control of the null quantile and, hence, of the statistical errors (type-I and type-II). Deb and Sen (2019) (Lemma 4.3) and Clémençon et al. (2009) (Theorem 2) derive *resp.* the exact and the asymptotic null distribution, that are independent of the underlying distributions. For the other methods, the corresponding test statistics depend on intrinsic unknown parameter(s) related to $G = H$, see *e.g.* Baringhaus and Franz (2004); Gretton et al. (2012a); Szekely and Rizzo (2004); Zhou et al. (2017). However, the explicit asymptotic distribution is obtained using the central limit theorem, see *e.g.* Bach et al. (2008) Theorems 1 and 3. Therefore, the estimation of the null requires data-driven estimation methods such as: bootstrap sampling, random permutation procedures, and moment matching methods, see *e.g.* Gretton et al. (2012a) Section 5. Such additional (sampling) procedures usually rely on large datasets to ensure sharp estimation. Methods circumvented this difficulty by modeling the two-sample problem as a classic semiparametric test. These boil down to testing a (set of) parameter(s). For instance, Zhou et al. (2017) and Baringhaus and Franz (2004) map the multivariate observations to the real line using a linear projection, leading to shift hypothesis testing.

The nonasymptotic bias of the proposed statistics is evaluated using concentration inequalities derived under, *e.g.*, boundedness or moment-based, assumptions. In Gretton et al. (2012a), tail bounds using Hoeffding (1963), depend on the class of bounded kernels (*e.g.* Th. 7 biased statistic, Th. 10 unbiased statistic, Th. 15 linear statistic). The majority of the statistics depend on the ambient representation of the observations or on directional information (*e.g.* linear projections Baringhaus and Franz (2004); Zhou et al. (2017)), leading to possibly biased statistics that cannot be ancillary (Basu (1959)), see *e.g.* Friedman and Rafsky (1979); Szekely and Rizzo (2004). In the context of complex data structures, and especially for observations sensitive to the data acquisition process, this is a main drawback for their use. The majority of the cited approaches guarantees asymptotic consistency of the tests but lacks advanced analysis to prove refined results, see *e.g.* Præstgaard (1995), Deb and Sen (2019) for asymptotic results (Theorems 4.3 and 4.4).

We lastly review computational properties. Either formulated as *plug-in* approaches or as an optimization problem, their implementation can be difficult and in particular for high dimensions (*e.g.*

large d, if $\mathscr{X} \subset \mathbb{R}^d$), also known as the *curse of dimensionality* phenomenon (*i.e.* when the required sample size for obtaining the convergence grows (exponentially) with d), see Devroye et al. (1996) Section 28.4. We also highlight that fundamental hyperparameters, such as the bandwidth of kernel functions, can require data-splitting procedures to learn the optimal one before being able to perform the procedure, see Gretton et al. (2007). This can be quite restrictive in practice, especially for small data samples. Heuristic choices for the optimal kernel bandwidth propose either using the empirical median or mean based on the interpoints distances. Gretton et al. (2012a) shows low efficiency in numerical experiments. Lastly, the computation of the test procedure can require techniques that are usually neither detailed nor provided in a companion algorithm. Online repositories coded by external contributors can be found while not maintaining usable versions. These are the significant limitations for using such advanced multivariate approaches. However, Baringhaus and Franz (2004); Deb and Sen (2019); Szekely and Rizzo (2004) are available in open access online libraries in the statistical software environment R.

We conclude by noticing that these *a priori* distinct families of tests are intimately related in nature. As studied in Sejdinovic et al. (2013), energy distances and kernel functions of a RKHS are related thanks to the simple equality $D(x,y) = (k(x,x) + k(y,y))/2 - k(x,y)$, $(x,y) \in \mathscr{X}^2$, for a distance D and a kernel k. Moreover, the metrics induced by these statistics are particular formulations of Wasserstein-based statistics, see Feydy et al. (2018); Ramdas et al. (2015). The methodology presented in this manuscript is different in nature and is inspired by the work of Clémençon et al. (2009) wherein the optimal statistic in learnt.

1.3 Rank processes

This section presents the chosen univariate definition of two-sample linear rank statistics. Then, we review multivariate extensions applied to the two-sample problem and detail their limitations. We refer to Appendix section B.1 for a comprehensive introduction in the univariate setting and in particular for classic methods (*e.g.* Hájek's projection, Hájek (1968)) and fundamental (asymptotic) properties under the null and the alternative hypothesis.

1.3.1 Univariate formulation

Historically, rank statistics were considered quite appealing thanks to their simplicity and fast computation for relatively small samples, formally starting with Spearman's rho test (Spearman (1904)) and later with Wilcoxon's two-sample test (Wilcoxon (1945)). Spearman motivated his statistic as a response to the traditional 'Gaussian assumption', as mainly rank methods *"reduce the "accidental errors""* (page 81, Spearman (1904)), compared to the ones based on the value of the observations themselves. Indeed, extremes, *i.e.*, observations far from the 'mean' behavior, do not 'weight' more in the computation of the statistic. On the contrary, these rare observations affect statistics which consider their values.

Let two independent random variables X, Y respectively drawn from G, H and valued in $\mathscr{X} \subset \mathbb{R}$. We consider two independent samples as follows. Let $X_1, \dots, X_n$, with $n \in \mathbb{N}^*$, *i.i.d.* observations drawn from G, and $Y_1, \dots, Y_m$, with $m \in \mathbb{N}^*$, *i.i.d.* drawn from H, such that $n/N \to p \in (0,1)$, with $N = n+m$. The parameter p is interpreted as the asymptotic proportion of the Xs among the pooled sample. We define univariate rank statistics based on the pooled sample of asymptotic mixture distribution equal to $F = pG + (1-p)H$. Under general assumptions on the underlying distributions, allowing for possible ties, we choose the definition of *upranks* (see van der Vaart (1998), page 173), as follows

$$\text{Rank}(t) = \sum_{i=1}^{n} \mathbb{I}\{X_i \leq t\} + \sum_{j=1}^{m} \mathbb{I}\{Y_j \leq t\}, \quad \text{for all} \quad t \in \mathscr{X} . \tag{1.3.1}$$

The ranks based on the pooled sample are, therefore, proportional to the empirical mixture distribution of F. By considering the empirical versions of the *c.d.f.* G and H, *i.e.* , $\widehat{G}_n(t) = (1/n)\sum_{i\leq n}\mathbb{I}\{X_i \leq t\}$ and $\widehat{H}_m(t) = (1/m)\sum_{j\leq m}\mathbb{I}\{Y_j \leq t\}$, for all $t \in \mathscr{X}$, its estimator is given by $\widehat{F}_N(t) = (n/N)\widehat{G}_n(t) + (m/N)\widehat{H}_m(t)$, since $n/N \to p$ as N tends to infinity. We obtain

$$\text{Rank}(t) = N\widehat{F}_N(t), \quad \text{for all} \quad t \in \mathscr{X} .$$

More generally, this definition allows for the use of empirical quantiles, copulas, *etc.* Two-sample linear R-statistics are constructed as a generic formulation encompassing various types of statistical tests. Solely the ranks of the Xs among the pooled samples are considered and tailored/weighted using a *score-generating function* $\phi : [0,1] \longrightarrow \mathbb{R}$, formally defined as one of the possible representations of the scores induced by the ranks, see Def. (B.1.3) and (B.1.4) for the main generating concepts.

Definition 2. (TWO-SAMPLE LINEAR RANK STATISTICS) *Let $\phi : [0,1] \to \mathbb{R}$ be a nondecreasing function. The two-sample linear rank statistics with 'score-generating function' $\phi(u)$ based on the random samples $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_m\}$ is given by*

$$\widehat{W}^{\phi}_{n,m} = \sum_{i=1}^{n} \phi\left(\frac{\text{Rank}(X_i)}{N+1}\right). \tag{1.3.2}$$

Fundamental results on two-sample R-statistics have been obtained thanks to H. Chernoff and I.R. Savage, and on the generalized version by J. Hájek. Chernoff and Savage (1958) provides an asymptotic analysis by writing the statistics as empirical measures using von Mises methods. Hájek and Sidák (1967) formalizes essential properties and examples of R-statistics. Dwass (1956) proposed for two-sample linear R-statistics a formulation thanks to *unbiased* (U-)statistics. He proved the asymptotic Gaussian distribution when all variables are identically distributed and, in particular, the study of the asymptotic power of the location test (Example 95). These results provide the foundations for studying linear rank statistics (Eq. (1.3.2)). In the context of two-sample testing, their fundamental property is their independence *w.r.t.* the underlying distributions under the null, see Appendix B, Lemma 87-(ii). It allows for the exact computation of critical values without any regularity assumptions on the underlying probabilities of the observations. We refer to Appendix section B for fundamental properties of such univariate statistics, both under the assumption of equality in distributions and under alternatives. Below, Figure 1.3 gathers some choices of ϕ leading to classic two-sample tests, which are in particular detailed in Appendix section B.

1.3.2 Multivariate extensions of R-statistics for the two-sample problem

This section introduces recent advances on multivariate generalizations of R-statistics applying to the two-sample problem. When considering the feature space $\mathscr{X}$ to be measurable and high-dimensional, defining rank statistics is far from straightforward due to the lack of natural order on $\mathscr{X}$. Even when considering $\mathscr{X}$ a subspace of $\mathbb{R}^d$, with $d \geq 2$, the adequate definition of these statistics is not clear. Hence, numerous methods motivated by the two-sample problem explored new concepts of ranks to circumvent this problem. They either rely on depth-based or spatial ranks or are usually designed for particular versions of the two-sample testing, or both, leading to (semi)parametric approaches. Lastly, $\mathscr{X}$ is supposed to be included in $\mathbb{R}^d$, with $d \geq 2$, and the references below are not exhaustive. We refer to the corpus of works led by M. Hallin for rank statistics applied to models for time series, see *e.g.* Hallin and Puri (1988, 1991); Hallin et al. (2020).

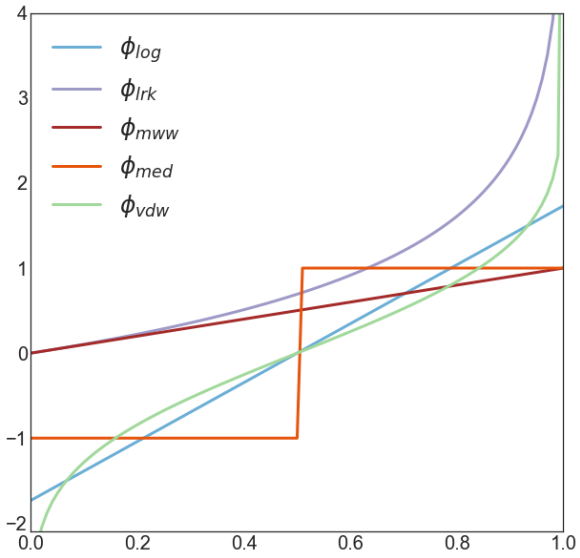


Figure 1.3. Curves of two-sample score-generating functions with the associated statistical test: Logistic test $\phi_{log}(u) = 2\sqrt{3}(u-1/2)$ in blue, Logrank test $\phi_{lrk}(u) = -\log(1-x)$ in purple, Mann- Whitney-Wilcoxon test $\phi_{mww}(u) = u$ in red, Median test $\phi_{med}(u) = \text{sgn}(u-1/2)$ in orange, Van der Waerden test $\phi_{vdw}(u) = \Phi^{-1}(u)$ in green, Φ being the normal quantile function.

Component-wise ranks. The former multivariate extension of ranks can be related to component-wise orderings, introduced by Hodges (1955) for bivariate signs and later studied in Puri and Sen (1993). Briefly, a univariate statistic is estimated based on the d-variate vector of the coordinate-wise ranks. The results are generally obtained under the independence of the coordinates, such that the vector of ranks is related to the d (univariate) marginals. Alternatively, some approaches considered semiparametric tests, such as the location one in Lung-Yut-Fong et al. (2015).

Data depth ranks. A more recent concept, inspired by results on probabilistic measure transportation, defines ranks *via* orderings induced by either center-outward distributions, quantile functions, or for instance, thanks to Monge-Kantorovitch ranks. These methods fall into the concept of *statistical depth*, see Mosler (2013). A *depth function* is a bounded measurable mapping from $\mathscr{Z}$ to $\mathbb{R}_+$ relative to a probability measure. It aims at defining a preorder for multivariate points of $\mathscr{Z}$. Also, it induces a notion of 'centrality' of the observations, such that the more the points are near the 'center' of the mass, and the more the depth function takes high values. Originally introduced in the seminal contribution of Tukey (1975), many alternatives have been developed since, *e.g.* Beirlant et al. (2020); Chaudhuri (1996); Chernozhukov et al. (2017); Deb and Sen (2019); Koshevoy and Mosler (1997); Liu (1990, 1995); Oja (1983); Vardi and Zhang (2000) and refer to Zuo and Serfling (2000) for a unified review on statistical depths. Liu and Singh (1993) proposed an extension of two-sample rank tests by two steps as follows. For a chosen depth, estimate the data-depth on a subset of the largest sample, then perform the univariate two-sample test on the values of the observations obtained by the empirical data-depth (of the first step). We finally refer to Hallin et al. (2021) for a comprehensive understanding of tests based on center-outward distributions.

Distance-based ranks. A rich literature was developed regarding (semi)parametric ranks, and interdirections based on the Mahalanobis statistical distance, from the one to the k-sample when considering elliptical distributions, see *e.g.* Hallin and Paindaveine (2002a,b, 2008) and for interdirections Um and Randles (1998). These were successfully applied to multiple types of homogeneity tests,

such as location, scale, and principal component analysis. Jurečková et al. (2010) introduced two definitions of ranks based on distances for the semiparametric location test, *i.e.*, under no assumptions on the probability class of distributions. They studied the ranks related to the distance of the observations to the origin and of the interpoint distances. Recently, and as already introduced in Section 1.2.2), an extension of rank statistics is obtained *via* optimal transport theory. The latter leads to learning *population metric rank maps* by optimizing the measure transportation from the unknown underlying distributions defined on the feature space to a reference distribution (usually the d-dimensional Halton sequence), see Deb and Sen (2019); Deb et al. (2021).

A particular corpus: the multivariate location model by H. Koul. The seminal works led by H. Koul and J. Jurečková provide an in-depth analysis of linear rank statistics for the multivariate location problem, under mild conditions regarding the score-generating function and the underlying distributions, see *e.g.* Gutenbrunner and Jurečková (1992); Koul (1970, 2002). These works present a straightforward multivariate extension of the semiparametric location model. The R-statistics are intimately linked to the linear regression as being computed on the corresponding regression error of one sample *w.r.t.* the other. In particular, R-statistics are studied in their most generic form *w.r.t.* the definition of the generalization of the scores (*i.e.* of $\phi(u)$), see Appendix B.1. Uniform asymptotic results on the location parameter and the score-generating function ϕ are provided under very mild conditions on the distributions. Moreover, these works studied the notion of contiguity for such multivariate generalizations.

1.3.3 Limitations

This section discusses on the properties of multivariate generalization concepts of rank statistics. Univariate rank statistics have fundamental properties, making them essential to the statistical literature. For instance, ranks lead to unbiased test statistics, achieving exact type-I error independent on the underlying distributions (Lemma 13.1, van der Vaart (1998)), and maximizing the power (uniform most powerful location test with Mann-Whitney-Wilcoxon statistic, see Rem. 14 and Lehmann and Romano (2005), Chapter 6.9), more details in B.1, B.2. Therefore, multivariate extensions are expected to guarantee similar properties. However, either typical asymptotic guarantees are obtained under mild conditions on the distributions, or refined results are proved at the price of parametric models, see Jurečková et al. (2010); Oja (2010). For instance, many fundamental contributions are obtained in semiparametric models, while deriving local analysis, particularly for families of elliptical distributions, see *e.g.* Hallin and Paindaveine (2002a,b, 2008).

Also, many generalization concepts rely on the ambient or local representations of the observations, or assume directional information. In particular, component-wise ranks are built on the chosen referential, see *e.g.* Hodges (1955). The corresponding assumptions being restrictive *w.r.t.* the underlying model, generic results thus lack and statistics are not distribution-free under both hypothesis. Similarly, depth-based ranks are inherent to the definition of the statistical depth itself, see Liu and Singh (1993). These tests are generally asymptotically distribution-free *w.r.t.* the null but fail in the sense of Basu (1959) (*i.e.* essential maximal ancillary). On the contrary, Chakraborty and Chaudhuri (2017); Chaudhuri (1996); Möttönen et al. (1997), *i.e.* spatial ranks, obtain statistics that are not distribution-free but are essential maximal ancillary.

As for the two-sample problem (Section 1.2), some concepts are implementable but at high computational costs. For instance, spatial ranks and, in particular, Oja medians face this difficulty. Ronkainen et al. (2003) proposed some deterministic and stochastic algorithms to reduce this complexity, *resp.* obtaining $\mathscr{O}(dN^d \log(N))$ and $\mathscr{O}(5^d \varepsilon^{-2})$ where ε is the radius of the confidence L_∞-ball. Later, Oja (2010) gathered methods for computing multivariate ranks in the statistical software

environment R. Deb and Sen (2019) can use assignment algorithms for computing the rank maps, that are at worst of order $\mathcal{O}(N^3)$. Clémençon et al. (2009) rely on bipartite ranking algorithms, that will be detailed in the next section. Also, few approaches are accompanied with numerical simulations, see *e.g.* Chakraborty and Chaudhuri (2017); Deb and Sen (2019), and incidentally rarely provided with accessible online code to reproduce the experiments.

To conclude, these concepts require a high level of statistical theory, and interesting properties are proved in the series of works. However, and to the best of our knowledge, none of the listed tests inherit from all the univariate properties (see section B.2), under mild conditions on the underlying probabilistic model. Most of the cited approaches rely on assumptions on the probabilistic model, often leading to its parametrization.

1.4 Learning-to-rank methods

Learning-to-rank methods are data-driven *ranking* approaches. The aim is to learn a relation order from a given set of observations according to their relevance/importance/preference, to predict the rank of any new set of instances. The learning task is formulated as a ranking one in a un/semi/-supervised setting. It finds interest in many research areas such as in Information Retrieval, Data Mining but also in recommendation systems (web search, mailing preference lists, *etc.*) and search engines. We first outline the broad context of ranking methods. Then, we detail the probabilistic formulation of pairwise models, known as *bipartite ranking*. State-of-the-art approaches for bipartite ranking are reviewed, as it is an essential model for the manuscript. Additionally, we highlight its dedicated section in Chap. 2 for a detailed review.

1.4.1 Context and formulation

In its most generic approach, the goal is to learn how to rank a set of observations, given a collection of queries, to minimize a statistical risk. It boils down to learning a scoring function $s(z)$ defined on a multidimensional feature space $\mathscr{Z}$ and valued in $\mathbb{R}$, such that one can rank any pair of instances: $a \preccurlyeq b$ *iff* $s(a) < s(b)$ where $<$ is the classic relation order in $\mathbb{R}$. For instance, in document retrieval, the goal is to learn an order for a set of documents *w.r.t.* their relevance, given a set of fields (queries). Importantly, these optimization problems consider the observations only through their *rank*, in the sense of the order statistics (Section 1.3). The fundamental question of such models is how to compare multivariate observations.

Multiple ranking methods exist to model various data structures. On the one hand, pointwise learning-to-rank methods address the problem of ranking *w.r.t.* the relevance of each labeled item. The loss function then evaluates the quality of the learned scoring function by comparing the prediction of each instance *w.r.t.* the ground truth. Ranking can therefore be modeled *via* classification, regression, or ordinal regression. On the other hand, pairwise and more broadly listwise methods, aim at formulating the loss function based on pair/list-wise comparison of items. The loss function measures the accuracy of the predicted pairs/lists of instances by a scoring function *w.r.t.* the ground truth. The associated algorithms are more complex, as they rely on at least pairwise relative comparison valued at all instances. We refer to Liu (2009) for a review of the models applied to Information Retrieval.

This manuscript focuses on a binary approach formulated as a pairwise learning-to-rank model with only one query, and known as *bipartite ranking* models. The observations are labeled with a binary variable, defined either as 'positive' or 'negative'. The goal is to learn their univariate mappings obtained thanks to an optimal *scoring function* $s(z)$. This function induces an order minimizing the bipartite ranking statistical loss. It recently has found interest in anomaly/novelty detection, where

the ranking is learned as to order the instances by the degree of their *abnormality*, see Clémençon and Jakubowicz (2013); Clémençon and Thomas (2018); Frery et al. (2017); Müller et al. (2013). We refer to Section 2.3, Chap. 2, for a review on those methods.

Probabilistic formulation of the bipartite ranking problem. Consider the input variable $\mathbf{Z}$ defined on the probability space $(\Omega, \mathscr{A}, \mathbb{P})$ and valued in the multivariate feature space $\mathscr{Z}$, associated to its binary label ζ valued in $\{-1, +1\}$. The heuristics of bipartite ranking can be reformulated as the comparison of two pairs of random variables $(\mathbf{Z}, \zeta)$ and $(\mathbf{Z}', \zeta')$ conditionally on the event $\{\zeta = 1, \zeta' = -1\}$ thanks to their value obtained by a *scoring function*. The optimal function s^* is learned from a class of candidates $\mathscr{S} = \{s : \mathscr{Z} \to \mathbb{R} \cup \{+\infty\}, \ s \text{ measurable}\}$, such that it minimizes the bipartite ranking risk defined by

$$L(s) = \mathbb{E}[\mathbb{I}\{s(\mathbf{Z}') > s(\mathbf{Z})\} \mid \zeta' = -1, \ \zeta = 1] + \frac{1}{2}\mathbb{P}\{s(\mathbf{Z}') = s(\mathbf{Z}) \mid \zeta' = -1, \ \zeta = 1\} , \tag{1.4.1}$$

where the ties are broken at random. Hence, s^* is defined by $L(s^*) = \inf_{\mathscr{S}} L =: L^*$. By considering the posterior probability $\eta(z) = \mathbb{P}\{\zeta = 1 \mid \mathbf{Z} = z\}$, it has been proved (Clémençon and Vayatis (2008), Proposition 2) that the set of optimal elements is given by

$$\mathscr{S}^* = \left\{ s \in \mathscr{S} \ \ s.t. \text{ for all } z, z' \text{ in } \mathscr{Z} : \ \ \eta(z) < \eta(z') \Rightarrow s^*(z) < s^*(z') \right\} . \tag{1.4.2}$$

We refer to Clémençon and Vayatis (2008) for related optimality results. The explicit bipartite ranking excess of risk for a scoring function $s(z)$ equals to

$$L(s) - L^* = \mathbb{E}[|\eta(\mathbf{Z}') - \eta(\mathbf{Z})| \mathbb{I}\{(s(\mathbf{Z}) - s(\mathbf{Z}'))(\eta(\mathbf{Z}) - \eta(\mathbf{Z}')) < 0\}] , \tag{1.4.3}$$

see Example 1 in Clémençon et al. (2008). This is the key for understanding many state-of-the-art methods as it will be discussed in the next section. In practice, the underlying distribution being unknown, we consider the statistical formulation based on *i.i.d.* random observations $\{(\mathbf{Z}_i, \zeta_i)_{i \leq N}\}$, with $N \in \mathbb{N}^*$. The goal of bipartite ranking is therefore to learn how to *score* any new sample $\mathbf{Z}_{N+1}, \ldots, \mathbf{Z}_{N+k}$ of unknown label, such that it minimizes the empirical counterpart of the expected loss function $L(s)$ when based on the training sample, defined by

$$\widehat{L}(s) = \frac{1}{nm} \sum_{\{i, \, \zeta_i = +1\}} \sum_{\{j, \, \zeta_j = -1\}} \left(\mathbb{I}\{s(\mathbf{Z}_j) > s(\mathbf{Z}_i)\} + \frac{1}{2}\mathbb{I}\{s(\mathbf{Z}_j) = s(\mathbf{Z}_i)\} \right) , \tag{1.4.4}$$

where $n = \sum_{i \leq N} \mathbb{I}\{\zeta_i = +1\}$ and $m = \sum_{i \leq N} \mathbb{I}\{\zeta_i = -1\}$. Ideally, the optimal scoring function reproduces the order induced by η and maximizes the scores of the 'positive' observations *w.r.t.* 'negative' ones. Refer to Menon and Williamson (2016) for a comprehensive review of the theoretical approaches to bipartite ranking and state-of-the-art algorithms.

1.4.2 Pairwise state-of-the-art approaches

This paragraph presents existing methods that aim at minimizing the empirical loss (Eq. (1.4.4)), or related formulations. We refer to Sections 2.2 and 2.3 for a detailed review on the approaches below.

A theoretical approach *via* Receiver Operating Characteristic (ROC) analysis. Formally introduced by Egan (1975), the ROC curve is a graph known as a gold standard tool for quantifying the dissimilarity between two populations. It is a Probability-Probability (P-P) plot of the True Positive Rate (TPR) *w.r.t.* the False Positive Rate (FPR) at all levels, see the Appendix section B.3 for its

properties related to univariate distribution functions. Therefore, it provides an excellent functional quality measure on the class $\mathscr{S}$ to discriminate between the scored 'positive' and 'negative' variables. More generally, the *Area Under the* ROC *Curve*, defined as the AUC, is intimately related to bipartite ranking as it exactly equals one minus the risk. We refer to Agarwal et al. (2005); Clémençon and Vayatis (2009b); Clémençon et al. (2008); Freund et al. (2003); Rudin (2006). Extensions to *multipartite ranking* are considered in Clémençon and Robbiano (2015); Clémençon et al. (2013b). In this sense, a series of works introduced different cost-sensitive functions to emphasize different manners of summarizing the ROC curve. First, a method extending tree-based methods was proposed by Clémençon and Vayatis (2009b); Clémençon et al. (2011), named `TreeRank`. Järvelin and Kekäläinen (2000) weighted the loss by the scores, defined as *Discounted Cumulative Gain* (DCG) factor. Boyd et al. (2012); Clémençon and Vayatis (2007) focused on only to the *best* instances, *i.e.*, those falling in the quantile of the whole mapped sample. Lastly, Agarwal (2011); Rudin (2006) proposed a smooth loss function named p/infinite-norm push, with $p > 0$, that 'pushes' the 'negative' instances far from the 'positive' ones. We incidentally highlight that ROC analysis is intimately related to rank statistics, as the empirical AUC is linearly proportional to the Mann-Whitney-Wilcoxon statistic.

Bipartite ranking risk as a pairwise classification loss. Most of learning-to-rank algorithms are built by transferring the pairwise loss of Eq. (1.4.1) to a univariate loss defined on pairs of *r.v.*. Briefly, letting $r_s : \mathscr{X} \times \mathscr{X} \to \{-1, 1\}$ a bivariate ranking rule depending on the scoring function s, such that $r_s(z, z') = 1 \quad \textit{iff} \quad s(z) \geq s(z')$, for all $(z, z') \in \mathscr{X}^2$, yields to considering the *r.v.* $((\mathbf{Z}, \mathbf{Z}'), (\zeta - \zeta')/2)$, as formulated in Clémençon et al. (2008) (Eq. (1.4.3)). Standard algorithms rely on this 'trick' and all the more on its extensions *via* surrogate margin losses. For instance, `RankBoost` (Freund et al. (2003)) is an extension of the `AdaBoost` (Freund and Schapire (1997)) leading to a bivariate exponential loss, that selects the *weak* ranker implying the largest decrease in the loss function. In Joachims (2006); Rakotomamonjy (2004), Support Vector Machines (SVM) are adapted to `RankSVM`, by minimizing the bipartite risk with surrogate hinge loss. Also, `RankNet` and `LambdaRankNet`, introduced by Burges et al. (2005), are adaptations of Neural Nets (NN) by optimizing the binary cross entropy loss with a modification of the backpropagation step. See also Narasimhan and Agarwal (2017) for SVM-based algorithm optimizing the partial AUC (defined in Appendix section B.3).

Bipartite ranking risk as a univariate loss. A series of works showed promising empirical results regarding the minimization of the bipartite ranking loss, obtained by using classic algorithms to minimize univariate losses (*e.g.* logistic, hinge, exponential losses). For instance, solutions of `AdaBoost`, logistic regression and even SVMs, show good empirical ranking performance, see *e.g.* Cortes and Mohri (2004); Rakotomamonjy (2004); Rudin and Schapire (2009). The heuristic relies on the class of optimal elements for both methods: in fact, they estimate the same oracle probability η, see Eq. (1.4.2). Hence, there are specific probabilistic frameworks for which optimal elements of univariate losses estimating class probabilities can be used as scoring functions for bipartite ranking. For theoretical analysis, we refer to the works of Agarwal (2014); Clémençon and Robbiano (2011); Narasimhan and Agarwal (2013); W. Kotlowski (2011).

Plug-in methods. As foreshadowed by Eq. (1.4.2), finding an optimal scoring function boils down to estimating the posterior probability or equivalently to the likelihood ratio. In fact, this corresponds to the so-called *plug-in* methods. Hence, if a *good* estimator of η is obtained, it suffices to use it as a scoring rule to map the observations, see Devore and Lorentz (1993). We refer to Clémencon and Vayatis (2009); Guedj and Robbiano (2018); Li et al. (2013) for approaches to ranking.

1.4.3 Limitations

This section details the significant computational limitations related to bipartite ranking and learning-to-rank methods. Refer to Section 2.2 for extensive details.

The minimization of the pairwise misranking error, or related surrogate forms, suffers from large scale implementation procedures (of order $\mathcal{O}(N^2)$), see Eq. (1.4.4). The nondifferentiability of the pairwise loss (indicator function) requires advanced analysis to guarantee that the (global/local) optimization is well defined. Techniques using alternative versions of the risk thanks to surrogate margin losses are detailed in Menon and Williamson (2016). This method led to fundamental results in binary classification, see *e.g.* Bartlett et al. (2006).

Pairwise classification models are at the heart of state-of-the-art ranking algorithms as an alternative formulation. While many interesting models have theoretical guarantees, we discuss on their construction. Indeed, it implicitly relies on the (strong) formulation of $r_s(z,z') = f_s(z-z')$, where $f_s : \mathscr{Z} \to \mathbb{R}$, see Eq. (1.4.3). When the feature space is $\mathscr{Z} \subset \mathbb{R}^d$, with $d \geq 2$, all algorithms suppose $\mathscr{S}$ parametric and composed of linear forms $z \in \mathscr{Z} \mapsto \langle \theta, z \rangle$, $\theta \in \Theta \subset \mathbb{R}^d$. Optimizing on $\mathscr{S}$ boils down to learning the optimal linear separation parameter θ and to univariate state-of-the-art algorithms, see *e.g.* `RankSVM`, `RankNN`, `RankBoost`. This drastically simplifies their implementation. Ailon and Mohri (2008) proposed algorithmic tricks to reduce the complexity for pairwise classification loss, from $\mathcal{O}(N^2)$ to $\mathcal{O}(N\log(N))$ ($\mathcal{O}(k\log(N+k))$ if top-$k \ll N$ instances considered). The complexity of linear `RankSVM` with L_1-loss reduced the quadratic complexity to at least $\mathcal{O}(N\tilde{n} + N\log(N))$, with $\tilde{n}$ the average number of non-zero features per observation, plus a loglinear term depending on the optimization algorithm, see *e.g.* Joachims (2006).

Despite such restrictive probabilistic frameworks and the remaining high complexity, those algorithms are empirically efficient, see *e.g.* Freund et al. (2003); Joachims (2006); Rakotomamonjy (2004). It leads to transferring ranking problems to classification or class probability estimation models. The main theorem of Balcan et al. (2007) proved that, given a binary classification loss, the obtained AUC $(1-L)$ is at most multiplied by 2. Narasimhan and Agarwal (2013) proved equivalence relations of the corresponding statistical risks under constraints, implying data-driven procedures (hence not distribution-free), see Diagram 2.2, Section 2.2. Agarwal (2014) formulated losses using *strongly proper* surrogate functions, *e.g.*, logistic and exponential losses.

Lastly, plug-in methods suffer from the curse of dimensionality due to the possible complex structure of $\mathscr{Z}$, see Section 1.2.3. Approaches may require, for instance, sparsity assumptions on the data or independence of the marginals of the conditional distributions, see *e.g.* Guedj and Robbiano (2018); Li et al. (2013). Finally, few methods gain leverage by providing grounded interpretability thanks to ROC analysis and incidentally to rank statistics, see *e.g.* Clémençon and Vayatis (2009b).

1.5 Contributions

This thesis aims to provide a generic framework for high-dimensional and nonparametric two-sample problems by studying linear rank processes. The main theoretical contributions are summarized in the first section, followed by two applications of such statistics in learning-to-rank problems, namely in bipartite and anomaly ranking. The last section, motivated by interdisciplinary research, outlines contributions on analyzing, modeling, and estimating the human postural control.

1.5.1 Towards a generic formulation for R-processes

Consider two independent *r.v.* $\mathbf{X}$ and $\mathbf{Y}$ *resp.* drawn from G and H, defined on a probability space $(\Omega, \mathscr{A}, \mathbb{P})$ and valued in the multivariate feature space $\mathscr{Z}$. Similarly to the univariate modeling and in the context of nonparametric statistics, let $p \in (0,1)$ and $N \geq 1/p$ such that we sample $n = \lfloor pN \rfloor$

and $m = \lceil (1-p)N \rceil = N - n$ two independent *i.i.d.* samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$, valued in $\mathscr{X}$ *resp.* from G and H.

As detailed in Section 1.3, defining rank statistics in high-dimensional spaces is far from straightforward. We propose a generalization of rank statistics based on bipartite ranking approaches for the whole manuscript. Precisely, by considering a class $\mathscr{S} = \{s : \mathscr{X} \to \mathbb{R} \cup \{+\infty\},\ s \text{ measurable}\}$ of scoring functions, the multivariate observations are analyzed through their univariate value obtained by a given $s(z)$. Two-sample *upranks* are defined by

$$\mathrm{Rank}(t) = \sum_{i=1}^{n} \mathbb{I}\{s(\mathbf{X}_i) \leq t\} + \sum_{j=1}^{m} \mathbb{I}\{s(\mathbf{Y}_j) \leq t\}\ , \quad \text{for all} \quad t \in \mathbb{R}\ , \tag{1.5.1}$$

where the subscript s is omitted for notation simplicity. This thesis studies theoretical properties of classes of the multivariate generalization of two-sample linear R-statistics (Eq. (2)), indexed on $\mathscr{S}$ and defined by

$$\widehat{W}_{n,m}^{\phi}(s) = \sum_{i=1}^{n} \phi\left(\frac{\mathrm{Rank}(s(\mathbf{X}_i))}{N+1}\right)\ . \tag{1.5.2}$$

This new definition allows for a generic form of the space $\mathscr{X}$ and provides interpretable ranks as they are valued in $\mathbb{R}$, inheriting the latter's statistical properties. This framework extends the works of S. Clémençon and N. Vayatis on ranking methods and rank-based estimators, see Clémençon and Vayatis (2007); Clémençon et al. (2008, 2009). Notice that choosing $\phi(u) = u$ recovers Mann-Whitney-Wilcoxon statistic, also know to be positively proportional to the *Area Under the* ROC *Curve* (AUC). Importantly, throughout this thesis we detail how this class of statistics can be interpreted as scalar performance criteria of learning-to-rank problems for learning the optimal $s(z)$. We show its relation to ROC analysis and how it reveals different characteristics of the underlying model depending on ϕ.

***R*-processes under scrutiny *via* PAC learning theory.** The heart of this thesis lies in the theoretical study of the collection of R-statistics as defined by Eq. (1.5.2), to provide guarantees for related applications. Importantly, finding the *optimal* scoring function such that the statistic converges to its continuous counterpart is analyzed thanks to PAC inequalities. For a given scoring function $s \in \mathscr{S}$, let F_s is the mixture *c.d.f.* of the two *r.v.* $s(\mathbf{X}) \sim G_s$ and $s(\mathbf{Y}) \sim H_s$ defined by $F_s = pG_s + (1-p)H_s$. We define the W_ϕ-ranking performance criterion by

$$W_\phi(s) = \mathbb{E}[(\phi \circ F_s)(s(\mathbf{X}))]\ , \tag{1.5.3}$$

The contributions focus on nonasymptotic analysis of the random and uniform fluctuations of $\{(1/n)\widehat{W}_{n,m}^{\phi}(s) - W_\phi(s)\}_{s \in \mathscr{S}_0}$, where $\mathscr{S}_0 \subset \mathscr{S}$. In particular, we are interested in guarantees on maximizers of the statistic over possibly infinite classes $\mathscr{S}_0$, *i.e.*,

$$\widehat{s} \in \arg\max_{s \in \mathscr{S}_0} \widehat{W}_{n,m}^{\phi}(s)\ . \tag{1.5.4}$$

However formulated as a typical Empirical Risk Minimization (ERM) problem, none of the classic tools can be directly applied, insofar as R-statistics are composed of sums of tailored non-*i.i.d.* correlated sums. It is essential to decompose and linearize the R-process due to its complex structure, to obtain a sum of empirical processes with nonasymptotic control of the error. Under polynomial control of the complexity of $\mathscr{S}_0$ (typically *Vapnik-Chervonenkis* (VC)-class), Proposition 3 below states this decomposition, for which the assumptions are detailed in Chapter 5.

Proposition 3 (Informal Proposition 53, Chapter 5)**.** *Suppose some assumptions on $\mathscr{S}_0 \subset \mathscr{S}$, ϕ and on the distributions of the two r.v.* $\mathbf{X}$, $\mathbf{Y}$. *The two-sample linear rank process* (5.3.5) *can be linearized/decomposed as follows. For all $s \in \mathscr{S}_0$,*

$$\widehat{W}^{\phi}_{n,m}(s) = n\widehat{W}_{\phi}(s) + \left(\widehat{V}^X_n(s) - \mathbb{E}\left[\widehat{V}^X_n(s)\right]\right) + \left(\widehat{V}^Y_m(s) - \mathbb{E}\left[\widehat{V}^Y_m(s)\right]\right) + \mathscr{R}_{n,m}(s)\ , \tag{1.5.5}$$

where

$$\begin{aligned}
\widehat{W}_{\phi}(s) &= \frac{1}{n}\sum_{i=1}^{n} (\phi \circ F_s)\left(s(\mathbf{X}_i)\right)\ , \\
\widehat{V}^X_n(s) &= \frac{n-1}{N+1}\sum_{i=1}^{n} \int_{s(\mathbf{X}_i)}^{+\infty} (\phi' \circ F_s)(u) dG_s(u)\ , \\
\widehat{V}^Y_m(s) &= \frac{n}{N+1}\sum_{j=1}^{m} \int_{s(\mathbf{Y}_j)}^{+\infty} (\phi' \circ F_s)(u) dG_s(u)\ .
\end{aligned}$$

For any $\delta \in (0,1)$,

$$\mathbb{P}\left\{\sup_{s\in\mathscr{S}_0} |\mathscr{R}_{n,m}(s)| < t\right\} \geq 1-\delta\ , \tag{1.5.6}$$

where t is of order $\log(1/\delta)$ and holds true for a particular range depending on p, ϕ, $\mathscr{S}_0$.

This fundamental result is the key for all the following analysis. It is based on multiple techniques, ranging from stochastic processes, such as U-processes, to chaining and decoupling methods for non-*i.i.d.* statistics, under minimal assumptions on the underlying probability distributions of the two samples, on the class $\mathscr{S}$ and on the function ϕ. Importantly, it required a new concentration uniform bound for U-processes based on two samples, that is stated and proved at length in its dedicated Chapter 4, Lemma 45 therein. It is now possible to guarantee nonasymptotic probability bounds of the *quality* of an empirical maximizer $\hat{s}$ (Eq. (1.5.4)) learnt within a class $\mathscr{S}_0 \subset \mathscr{S}$, yielding to the following generalization bound and proved in Corollary 56, Chapter 5 therein.

Corollary 4 (Informal Corollary 56, Chapter 5)**.** *Let $\hat{s}$ be an empirical maximizer of the W_ϕ-ranking performance over the class $\mathscr{S}_0$. Under the assumptions of Proposition 3, for any $\delta \in (0,1)$, we have with probability at least $1-\delta$:*

$$W^*_{\phi} - W_{\phi}(\hat{s}) \leq 2C_2\sqrt{\frac{\log(C_1/\delta)}{pN}} + \left(W^*_{\phi} - \sup_{s\in\mathscr{S}_0} W_{\phi}(s)\right), \tag{1.5.7}$$

where W^*_ϕ is the oracle measure, valued at scoring functions of strictly increasing transforms of the likelihood ratio dG/dH. The inequality holds true for a particular range of δ depending on p, where the constants C_1, C_2 depend on ϕ, $\mathscr{S}_0$.

The result above establishes that the maximizers $\hat{s}$ (Eq. (1.5.4)) achieve classic learning rate bound of order $O_{\mathbb{P}}(1/\sqrt{N})$ when based on training datasets of size N, just like in standard classification, see *e.g.* Devroye et al. (1996). Additional contributions are detailed in the main corpus, related to guarantees on model selection procedures penalized by the class complexity, for which generalization bound are stated in Corollary 57 Chapter 5. Lastly, a smoothed counterpart is introduced using kernel density estimation methods to ensure the local concavity and differentiability of the statistic. It results to similar guarantees of Corollary 4 with an additional term due to the regularization, see Proposition 58 Chapter 5.

1.5.2 *R*-processes applied to two-sample homogeneity testing

Consider the framework introduced previously. We propose a generic two-stage method applied to the high-dimensional two-sample problem (Section 1.2) using the *R*-statistics $\widehat{W}^{\phi}_{n,m}$, when indexed by a class of scoring functions $\mathscr{S}$ (see Eq. (1.5.2)). As foreshadowed in Eq. (1.5.1), for a given $s : \mathscr{X} \to \mathbb{R} \cup \{+\infty\}$, we define two-sample ranks by comparing the mapped values of the observations by $s(z)$. The obtained univariate random samples are $s(\mathbf{X}_1), \ldots, s(\mathbf{X}_n)$ and $s(\mathbf{Y}_1), \ldots, s(\mathbf{Y}_m)$. At level of test $\alpha \in (0,1)$, for a given score-generating function $\phi(u)$ and a scoring function $s(z)$, the test statistic based on the two samples $\mathscr{D}_{n,m} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\} \cup \{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$ is defined by

$$\Phi^{\phi}_{\alpha}(\mathscr{D}_{n,m}(s)) = \mathbb{I}\left\{\widehat{W}^{\phi}_{n,m}(s) > q^{\phi}_{n,\,m}(\alpha)\right\}, \tag{1.5.8}$$

where $q^{\phi}_{n,\,m}(\alpha)$ is the $(1-\alpha)$-quantile of the null distribution of the statistic $\widehat{W}^{\phi}_{n,m}$ for fixed sample sizes n, m. Notice that it is independent of s. The two-sample problem is reformulated as follows

$$\mathscr{H}_0 : \; W^*_{\phi} = \int_0^1 \phi(u)du \quad \textit{versus} \quad \mathscr{H}_1 : \; W^*_{\phi} > \int_0^1 \phi(u)du \,, \tag{1.5.9}$$

where W^*_{ϕ} is obtained when valued at the oracle class $\mathscr{S}^*$ of strictly nondecreasing transforms of the likelihood ratio dG/dH.

In this line, a two-stage procedure is based on: 1. *bipartite ranking:* to learn the optimal scoring function on the first half of each sample, and 2. *two-sample homogeneity test:* using the optimal scoring function obtained at 1. to map the second halves, perform the hypothesis test of (1.5.9) using the test statistic (1.5.8). Fig. 1.4 summarizes the procedure.

Theoretical guarantees of *Step* 1. This relies on the intrinsic relation between the minimization of bipartite ranking loss formulations and the maximization of *R*-statistics as scalar criteria for the class of scoring functions. In fact, under regularity conditions, *Step 1* aims to learn strictly monotonous transforms of the likelihood ratio of the underlying distributions, guaranteed thanks to Chapter 5. Hence it ignores the *curse of dimensionality* and possible model bias issues while satisfying the rank statistics-related properties. It is asymptotically consistent and also achieves competitive generalization bounds of order $\mathscr{O}_{\mathbb{P}}(N^{-1/2})$.

Theoretical guarantees of *Step* 2. As outlined in Section 1.2, few state-of-the-art methods are able to derive proper theoretical guarantees on multivariate statistics, and especially nonasymptotic ones. We first prove nonasymptotic distribution-free control of the *R*-statistics, by means of concentration bounds under both hypothesis, gathered in Proposition 5. The asymptotic distributions are also detailed.

Proposition 5 (Informal Propositions 71 and 72, Chapter 6)**.** *Suppose some assumptions on ϕ, G, H and consider $s \in \mathscr{S}_0$ fixed, e.g. the optimal element of the first step of the procedure. Then under $\mathscr{H}_1$, for all $t > 0$, $N \geq 2$:*

$$\mathbb{P}_{\mathscr{H}_1}\left\{|\widehat{W}^{\phi}_{n,m}(s) - W_{\phi}(s)| > t\right\} \leq 20e^{-CNt^2/8}\,, \tag{1.5.11}$$

where C depends on ϕ, p, and for a particular range on t depending on p, ϕ, N. Under the null hypothesis $\mathscr{H}_0$, the following inequality holds true for all $t > 0$:

$$\mathbb{P}_{\mathscr{H}_0}\left\{|\widehat{W}^{\phi}_{n,m} - W^0_{\phi}| > t\right\} \leq 2e^{-2pN(t-\Delta_N)^2}\,, \tag{1.5.12}$$

where $\Delta_N = |(1/N)\sum_{i\leq N}\phi(i/(N+1)) - W^0_{\phi}|$ and $W^0_{\phi} = \int_0^1 \phi$.

RANKING-BASED TWO-SAMPLE RANK TESTS

Input. Two independent and *i.i.d.* samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$ of sizes n, $m \geq 2$ and valued in $\mathscr{Z}$ - subsample sizes $n' < n$ and $m' < m$ - bipartite ranking $\mathscr{A}$ algorithm operating on the class $\mathscr{S}_0$ of scoring functions on $\mathscr{Z}$ - univariate two-sample rank test Φ_α of level $\alpha \in (0,1)$.

Two-split trick. Divide each of the original samples into two subsamples:

$$\{\mathbf{X}_1, \ldots, \mathbf{X}_{n'}\} \cup \{\mathbf{X}_{1+n'}, \ldots, \mathbf{X}_n\} \text{ and } \{\mathbf{Y}_1, \ldots, \mathbf{Y}_{m'}\} \cup \{\mathbf{Y}_{1+m'}, \ldots, \mathbf{Y}_m\}$$

1. **Bipartite ranking.** Learn the maximizer $\hat{s}$ by $\mathscr{A}$, based on training data $\{\mathbf{X}_1, \ldots, \mathbf{X}_{n'}\}$, $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_{m'}\}$.
2. **Univariate rank test.** Form the univariate samples

$$\{\hat{s}(\mathbf{X}_{1+n'}), \ldots, \hat{s}(\mathbf{X}_n)\} \text{ and } \{\hat{s}(\mathbf{Y}_{1+m'}), \ldots, \hat{s}(\mathbf{Y}_m)\} \, .$$

Compute the α-level test statistic

$$\Phi_\alpha^\phi \left(\{\hat{s}(\mathbf{X}_{1+n'}), \ldots, \hat{s}(\mathbf{X}_n)\} \cup \{\hat{s}(\mathbf{Y}_{1+m'}), \ldots, \hat{s}(\mathbf{Y}_m)\} \right) \, , \tag{1.5.10}$$

depending on $\{\text{Rank}(\hat{s}(\mathbf{X}_{1+n'}), \ldots, \text{Rank}(\hat{s}(\mathbf{X}_n))\}$, where $\text{Rank}(t) = \sum_{i=1+n'}^{n} \mathbb{I}\{\hat{s}(\mathbf{X}_i) \leq t\} + \sum_{i=1+m'}^{m} \mathbb{I}\{\hat{s}(\mathbf{Y}_j) \leq t)\}$.

Figure 1.4. Ranking-based two-sample rank test procedure.

The proofs rely on classic concentration inequalities as well as the ones obtained in Chapter 4, applied to the terms inherited by the decomposition of Proposition 3 (see Prop. 53, Chapter 5). Additionally, an estimator of the power is provided based on a Monte-Carlo sampling scheme. The asymptotic distributions of the (studentized) statistics are stated in Propositions 73, 74 (under $\mathscr{H}_1$) and 75, 76 (under $\mathscr{H}_0$), Chapter 6. Importantly, the null distribution depends only on ϕ and p. Lastly, in order to leverage on the score-generating function, we propose a practical procedure, as to choose the *optimal* ϕ in a minimax sense.

Numerical experiments. A series of numerical experiments are conducted for multiple probabilistic models, testing a range of state-of-the-art bipartite ranking algorithms as listed in Section 1.4.2. Additionally, comparisons to state-of-the-art multivariate tests are automatically presented, in particular to Friedman and Rafsky (1979); Gretton et al. (2012a); Szekely and Rizzo (2004). They are gathered in Chapter 7 and are accompanied with open access online `Python` codes available at `https://github.com/MyrtoLimnios`.

1.5.3 *R*-processes applied to learning-to-rank problems

The generic framework for two-sample linear *R*-processes is applied to two models inherited from the machine learning community: bipartite ranking and anomaly detection problems. Indeed, the continuous counterpart W_ϕ is interpreted as scalar performance criterion, that summarizes fundamental functional losses in ranking problems. For both topics, we provide an algorithmic procedure.

Bipartite ranking. Based on a multivariate *i.i.d.* sample with random binary labels, the goal is to learn a scoring function such that one can rank any new observation of unknown label and with minimum ranking error. We justify the *almost* equivalence to the two-sample formulation in Section 2.2 (Chap. 2), to be able to consider the proposed R-statistic (1.5.2). The smooth version of the R-statistic proposed in Chapter 5 allows, in particular, to implement a deterministic gradient ascent algorithm. The goal is to maximize an empirical version of the smoothed statistic by introducing a second-order Parzen-Rosenblatt kernel, K of bandwidth h, such that

$$\widehat{W}^{\phi}_{n,m,h}(s) = \sum_{i=1}^{n} (\phi \circ \widehat{F}_{s,N,h})(s(\mathbf{X}_i)) \ , \tag{1.5.13}$$

where $\widehat{F}_{s,N,h}$ is the empirical mixture distribution of the pooled sample, regularized with the kernel K of bandwidth h. The procedure is summarized in Algorithm 1, where the class $\mathscr{S}_0 = \{s_\theta : \mathscr{X} \to \mathbb{R}, \ \theta \in \Theta\}$ is supposed parametric *w.r.t.* $\Theta \subset \mathbb{R}^d$, and numerical results are gathered in Chapter 7 that rely on the guarantees of Chapter 5. The algorithm is coded in `Python` and is accessible at the open access online repository at `https://github.com/MyrtoLimnios`.

Algorithm 1: Gradient Ascent for W-ranking performance maximization

Data: Independent *i.i.d.* samples $\{\mathbf{X}_i\}_{i\leq n}$ and $\{\mathbf{Y}_j\}_{j\leq m}$.
Input: Score-generating function ϕ, kernel K, bandwidth $h > 0$, number of iterations $T \geq 1$, step size $\eta > 0$.
Result: Scoring rule $s_{\widehat{\theta}_{n,m}}(z)$.

Choose the initial point $\theta^{(0)}$ in Θ;
for $t = 0, \ldots, T-1$ **do**
  compute the gradient estimate $\nabla_\theta \left(\widehat{W}^{\phi}_{n,m,h}(s_{\theta^{(t)}}) \right)$;
  update the parameter $\theta^{(t+1)} = \theta^{(t)} + \eta \nabla_\theta \left(\widehat{W}^{\phi}_{n,m,h}(s_{\theta^{(t)}}) \right)$;
end
Set $\widehat{\theta}_{n,m} = \theta^{(T)}$.

Anomaly ranking. We propose a ranking approach to anomaly detection, where R-statistics are used as scalar criteria to learn a scoring rule for ordering the instances by *degree* of abnormality. This method aims to bridge the gap between one-sample unsupervised anomaly detection and two-sample rank-based discrepancy measure of probability distributions. This formulation is theoretically interesting as it is framed for discriminating between two *a priori* different distributions independently on the *shape* of the perturbation. However, traditional methods are based on criteria minimizing a local/global spatial risk function, which can lead to ranking the observations based on an accuracy index independent of a ranking criterion.

The statistical learning framework slightly differs from previously. Let $p \in (0,1)$, we assume that $N \geq 2$ observations are available: $n = \lfloor pN \rfloor$ 'normal' *i.i.d.* observations $X_1, \ldots, X_n$ taking their values in $[0,1]^d$ for simplicity drawn from $F(dx) = f(x)\lambda(dx)$. Let $m = N - n$, the *i.i.d.* sample $U_1, \ldots, U_m$ drawn from the uniform distribution $\mathscr{U}_d$ on the hypercube $[0,1]^d$, independent from the X's. Hence, p represents the 'theoretical' proportion of 'normal' observations among the pooled sample. Similarly to bipartite ranking, the goal is to estimate the optimal scoring function s, such that the empirical W_ϕ-ranking criterion is maximized. In fact, the latter is shown to be intimately related to a functional criterion introduced in Clémençon and Jakubowicz (2013), known as *Mass-Volume* (MV) curves.

The MV curve of a scoring function $s \in \mathscr{S}_0 \subset \mathscr{S}$ can be defined as the parametric curve

$$\mathrm{MV_s : t \in \mathbb{R} \mapsto (1 - F_s(t), 1 - \lambda_s(t)) \in [0,1]^2 \ ,}$$

where $F_s(t) = \mathbb{P}\{s(\mathbf{Z}) \leq t\}$ and $\lambda_s(t) = \lambda(\{z \in \mathscr{Z},\ s(z) \leq t\})$, for all $t \in \mathbb{R}$. A two-stage procedure is proposed and summarized in Fig. 1.5, based on: 1. *maximization of the empirical W_ϕ-ranking performance criterion* based on $\{X_1, \ldots, X_n\}$ and $\{U_1, \ldots, U_m\}$; 2. *ranking the anomalies* of a new sample $\{X_1^t, \ldots, X_{n_t}^t\}$. Numerical experiments are provided in Chapter 8, where the algorithm approximates *Step 1*. It is based on a Neural Network with a binary cross entropy loss, penalized by the W_ϕ-criterion.

ANOMALY RANKING PROCEDURE

Input. Consider the 'normal' *i.i.d.* random sample $\{X_1, \ldots, X_n\}$ of unknown probability probability function $F(dt)$ defined on $\mathscr{Z} \subset [0,1]^d$, $d \geq 2$ - new random sample $\{X_1^t, \ldots, X_{n_t}^t\}$ *a priori* drawn from $F(dt)$, $n_t \in \mathbb{N}^*$, independent of the X's - bipartite ranking $\mathscr{A}$ algorithm operating on the class $\mathscr{S}_0$ of scoring functions on $\mathscr{Z}$ - ϕ a score-generating function. Set $n_{lowest} \in \mathbb{N}^*$.

1. **Maximizing the W_ϕ-ranking performance criterion.**
   (i) Generate a *i.i.d.* random sample $\{U_1, \ldots, U_m\}$ from $\mathscr{U}_d([0,1]^d)$, independent of $\{X_1, \ldots, X_n\}$,
   (ii) Output the optimal empirical scoring rule $\hat{s}_{n,m} = \arg\max_{s \in \mathscr{S}_0} \widehat{W}_{n,m}^{\phi}(s)$ based on the two samples $\{X_1, \ldots, X_n\}$ and $\{U_1, \ldots, U_m\}$.
2. **Ranking anomalies.**
   (i) Compute the empirical scores of the test sample $\{\hat{s}_{n,m}(X_1^t), \ldots, \hat{s}_{n,m}(X_{n_t}^t)\}$,
   (ii) Define as anomalous the n_{lowest} observations of lowest empirical scores among the sequence $\hat{s}_{n,m}(X_i^t)$, $i \leq n_t$.

Result. Output the set of anomalous observations with their corresponding ranks.

Figure 1.5. Two-stage procedure for learning to rank anomalies.

1.5.4 Analyzing, modeling and estimating the postural control

We present two main contributions related to biomedical studies and in particular to the analysis of the postural control, introduced in Section 1.1. Briefly, postural control is usually measured using sensorimotor platforms, registering the temporal variation of the *Center of Pressure* (CoP) displacement (statokinesigram) of the patient during a short timescale. The obtained measurements are longitudinal and two-dimensional in space, refer to Figures 1.1 (protocol) and 1.2 (examples of statokinesigrams).

An interpretable algorithm for the two-sample problem applied to biomedical studies. We explore an easy-to-use interpretable algorithm for performing the two-sample problem on clinical populations. Precisely, the observations are retrieved from real measurements of statokinesigrams for two classes of patients: *Fallers* (frail population) and *non Fallers* ('control' population). The obtained observations are of complex structures that are difficultly processed and analyzed by classic methods.

For the statistical comparison of two cohorts, the proposed method is inherited from Chapter 6 but with different learning procedure of Fig. 1.4. Biomedical studies are typically characterized by few patients, *i.e.*, small sample sizes. Thus, procedures based on data-splitting do not necessarily converge. We propose a cross-validation sampling procedure, associated with a learning algorithm in *Step 1* based on a random forest combined with an out-of-bag algorithm (leave-one-out). For *Step 2.*, we choose $\phi(u) = u$, leading to Mann-Whitney-Wilcoxon test statistic.

Interesting empirical results are obtained, competing as well state-of-the-art methods. In particular, it shows similar results to the classic *Maximum Mean Discrepancy* (MMD) test, see Gretton et al. (2012a). This method also provides a nice interpretation that is valuable for the biomedical community, such as feature importance extraction, robustness *w.r.t.* small sample sizes and imbalanced samples. Lastly, it reveals findings (similarly to the MMD) that were not obtained when using classic multiple testing procedures with corrections, known to control the type-I error (*e.g.* Bonferroni, Holmes, Sidàk corrections, see Hochberg (1988); Hommel (1988)). It is, therefore, a first step towards communicating the importance of using accurate multivariate methods for such complex data.

A stochastic model to understand postural control. This second contribution provides a generative model for predicting the temporal evolution of the CoP. Precisely, we introduce the Local Recall model, where the CoP is assumed to be the solution of a modified Langevin stochastic differential equation (SDE), ruled by the trajectory of the *Center of Mass* (CoM). This is quite a new formulation in this class of models, wherein CoP and CoM are temporally correlated. A significant difficulty is to settle/choose the discretization paradigm related to the continuous SDE model. Indeed, it explores mathematical SDE tools ensuring the stability of the system, as well as various biomechanical interpretations of such choices. The procedure obtained is two-fold: (i) based on the recorded statokinesigrams, estimate the parameters of the discrete system, and (ii), generate the learned trajectories for the CoP using the estimated parameters. A series of numerical experiments are provided, for which empirical results show a low statistical error of the generated trajectories in the sense of the least square error measure.

1.6 Additional information

1.6.1 Manuscript organization

The manuscript is composed of three parts that are briefly described in the sequel with, eventually, the associated publication(s).

Existing two-sample problems and mathematical background. This part gathers and formulates the main concepts of the thesis. It is divided into two chapters:

- *Chapter 2: Two-sample problems.* First, the multivariate and nonparametric two-sample problem is reviewed. Then, two learning-to-rank models are detailed, namely the bipartite and anomaly ranking problems. All are formulated in their generic form, while reviewing in details the main results and state-of-the-art methods.

- *Chapter 3: Some concentration results.* Motivated by Empirical Risk Minimization, the construction of classical concentration inequalities are outlined. This leads to bounding statistics and more importantly collection of them, referred-to as empirical processes (if of order 1), U-processes (of greater order).

Contributions related to *R*-processes. This second part gathers the core of the thesis, rendering the analysis of the generalized version of *R*-processes and its application to the two-sample problem. The numerical experiments on synthetic data are gathered in the last Chapter.

- *Chapter 4: A concentration inequality for U-processes.* This introductory chapter provides a new concentration bound for a particular degenerate two-sample U-process, when indexed by a class of kernels of 'controllable' complexity. It is a new result in the literature that is required for the following chapter and hence was published as part of the article [1].
- *Chapter 5: Concentration inequalities for two-sample R-processes.* Results on *R*-processes are proved and motivated by the bipartite ranking modeling. It corresponds to the publication [1].
- *Chapter 6: Two-sample homogeneity testing.* A generic formulation for the two-sample problem based on *R*-statistics is proposed, optimized thanks to learning-to-rank algorithms. We state a two-stage procedure with proved theoretical guarantees. This is still a working paper.
- *Chapter 7: Numerical experiments.* This section gathers numerical experiments based on synthetic data in order to test our proposed rank-based criteria in two contexts: bipartite ranking and two-sample testing. All the details on the codes are additionally detailed. It gathers the numerical results of [1] (Chapter 6) and [2] (Chapter 8). The algorithms are coded in `Python` and are accessible at the open access online repository at `https://github.com/MyrtoLimnios`.

Applications. This last part focuses on three applied contributions. While the first is related to a learning-to-rank model, the following two result from the interdisciplinary research at the Centre Borelli and related to the analysis and modeling of the postural control.

- *Chapter 8: Learning to rank anomalies with two-sample linear R-statistics.* We derive a methodology for learning to rank observations by their *degree* of abnormality. The associated publication is [3].
- *Chapter 9: Two-sample testing applied to biomedical studies.* A two-sample homogeneity testing method for biomedical applications is detailed, fitted to maximize a particular version of the proposed *R*-statistics. This algorithm is applied to the statistical comparison of two clinical populations, composed of measurements retrieved from statokinesigrams. The associated publications/communications are [4-5].
- *Chapter 10: A generative model for the postural control.* A model is proposed to generate the temporal evolution of the center of pressure when modeled to be temporally correlated to the center of mass. It is based on the stochastic Langevin model. The associated publication is [6].

Appendix. The Appendix gathers three chapters as follows.

- *Appendix A: Generalized two-sample R-processes and efficient two-sample tests.* This section is related to the study of *R*-statistics when indexed by a class of score-generating functions ϕ. In the continuity of Chapter 5 and inspired by the work of H. Koul (see Section 1.3.2), we study of *R*-processes under mild assumptions on the score-generating function. Then, for the two-sample problem, an additional procedure is proposed, wherein *Step 1.* of Fig. 1.4 is replaced by the exact maximization of the *R*-statistic. We use the Algorithm 1 and some numerical experiments are provided. Also, a adaptive approach for choosing the 'best' score-generating function is detailed.

- *Appendix B: Univariate framework and state-of-the-art.* This section develops fundamental results and examples on the following univariate problems: two-sample rank statistics, the two-sample problem and ROC analysis.

- *Appendix C: Additional material.* Some facts on scientific research based on statistics are raised, especially for replicable research. Lastly, the general introduction in French is outlined.

1.6.2 References and online material

The following list gathers the publications and a working paper that are considered in this manuscript. The majority of articles were published in scientific journals, whereas the remaining two in a conference and a workshop.

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

[2] *A bipartite approach to the two-sample problem* S. Clémençon, M. Limnios*, N. Vayatis. Working paper. * *Corresponding author*

[3] *Learning to rank anomalies: scalar performance criteria and maximization of two-sample rank statistics*. M. Limnios, N. Noiry, S. Clémençon. In *Proceedings of the Third International Workshop on Learning with Imbalanced Domains: Theory and Applications*, volume 154 of Proceedings of Machine Learning Research, pp. 63–75, Sep. 2021.

[4] *Revealing posturographic features associated with the risk of falling in patients with Parkinsonian syndromes via machine learning.* I. Bargiotas, A. Kalogeratos, M. Limnios, P-P. Vidal, D. Ricard, N. Vayatis. PLOS ONE 16(2): e0246790, Feb. 2021.

[5] *Multivariate two-sample hypothesis testing through AUC maximization for biomedical applications.* I. Bargiotas, A. Kalogeratos, M. Limnios, P-P. Vidal, D. Ricard, N. Vayatis. SETN 2020: 11th Hellenic Conference on Artificial Intelligence, pp 56–59, Sep. 2020.

[6] *A Langevin-based model with moving posturographic target to quantify postural control.* A. Nicolai, M. Limnios, A. Trouve, J. Audiffren. IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 29, pp. 478-487, Feb. 2021.

Additionally, during the outset of the pandemic outbreak, I was mainly involved in a research project mapping published propagation models of the epidemic. It led to a project of collecting and reviewing mainly epidemiological articles. We gathered our results into an open and online repository that includes a tabular sheet, a `Kibana` interface and a markdown document. The aim is to facilitate the identification of models by proposing a cartography of the approaches proposed from February to May 2020. Moreover, the joint pre-print [7] was published that focuses on: (i) the epidemic propagation models, (ii) the modeling of intervention strategies, (iii) the models and estimation procedures of the epidemic parameters and (iv) the characteristics of the data used.

[7] *Epidemic models for COVID-19 during the first wave from February to May 2020: a methodological review*, M. Garin*, M. Limnios*, A. Nicolai*, I. Bargiotas, O. Boulant, S. E. Chick, A. Dib, M. Fekom, A. Kalogeratos, C. Labourdette, A. Ovchinnikov, R. Porcher, C. Pouchol, T. Evgeniou, N. Vayatis. *Preprint, arXiv:2109.01450*, 27 pages, Sep 2021. * *Equal contribution*

Lastly, as we believe in the importance of replicable, easy-to-use and open-access tools, the respective algorithms of the aforementioned projects are gathered into the github repository https://github.com/MyrtoLimnios. Therefore, the Python codes used in the articles and the online repository joint to the review [7] are easily accessible.

Part I

Problems with Two Samples and Concentration Inequalities

2 | Two-sample Problems

Abstract. This preliminary chapter outlines three two-sample problems studied in the manuscript. The first section presents the multivariate and nonparametric homogeneity hypothesis testing, known as the two-sample problem. Precisely, state-of-the-art methods for the nonparametric statistical comparison of two independent random samples are outlined when supposed to be valued in a multivariate space. The following sections aim to provide an insight into *ranking* problems. These are formulated as optimization models wherein the observations resulting from an experiment are considered solely through their *ranks* (in the sense of the order statistics), either in a *supervised* setting or in a *unsupervised* one. In particular, the focus is on two problems: (i) the bipartite ranking and (ii) the anomaly ranking. While the former gathers rich literature for the past decades, the latter is a recent formulation in anomaly/novelty detection. For multivariate observations, the definition of the *ranks* is not unique. We present one of the existing modelings that introduces a *scoring function* inducing an order by mapping the observations from the feature space to the real line. In particular, *Receiver Operating Characteristic* (ROC) analysis is introduced, insofar as it induces a *quality* criterion for the scoring functions, that is either functional (if considering the ROC curve) or scalar (else summaries of the ROC curve). Additionally, we present their two-sample formulation that will be considered throughout the manuscript.

Contents

2.1 Homogeneity testing

This section highlights recent methods developed for the two-sample problem in nonparametric and multivariate frameworks. While a rich literature for generalizing (semi)-parametric tests exists, particularly for the location and scale tests, few works propose a generic nonparametric approach. The majority of the proposed approaches estimate a metric between the underlying distributions of the two samples. Statistics are usually modeled in particular settings, *e.g.*, for precise data structures or families of distributions. We refer to alternative models based on: nearest-neighbors tests Henze (1988); Schilling (1986), matching/assignment Mukherjee et al. (2020), permutation tests Hall and Tajvidi (2002), classifier Lopez-Paz and Oquab (2016), random projections Lopes et al. (2011); Srivastava et al. (2016), random forest Hediger et al. (2021), sparse mixture model Arias-Castro and Wang (2017), differential privacy Couch et al. (2019); Lam-Weil et al. (2020); Si et al. (2021). We refer to the review of Bhattacharya (2019); Lovato et al. (2020) for tests applied to graph structures or formulated via multivariate data-depths. Lastly, a related problem in computer science literature refers to two-sample problem as property testing, see for instance Goldreich et al. (1998); Rubinfeld and Sudan (1996).

First, we recall the generic formulation of the two-sample problem. We then review generalizations of classic univariate statistics and metric-based approaches, estimating a distance between the probability measures (or related) of the two samples. The heuristic is that the null hypothesis $\mathscr{H}_0$ is equivalent to obtaining the chosen distance equal to zero. We additionally point the main properties and limitations.

2.1.1 Formulation

Consider two independent random variables $\mathbf{X}$ and $\mathbf{Y}$, defined on a probability space and valued in the (same) multivariate measurable space $\mathscr{X}$, of unknown continuous distribution functions G and H. For a fixed level $\alpha \in (0,1)$, the two-sample problem corresponds to testing the two hypothesis below:

$$\mathscr{H}_0 : \; G = H \text{ against the alternative } \mathscr{H}_1 : \; G \neq H \; . \tag{2.1.1}$$

Also known as homogeneity testing, many classic statistical problems can be related to this generic formulation. See Darling (1957) for the univariate goodness-of-fit testing and Friedman (2004) for the multivariate model, Spearman (1904) for independence testing, and Wilcoxon (1945) for pairwise testing. In practice, and especially for nonparametric settings, we consider independent copies of the

r.v. as the underlying (classes of) distributions are unknown. Let $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$, with n, $m \in \mathbb{N}^*$, two independent *i.i.d.* samples drawn from G and H and valued in the (same) measurable space $\mathscr{Z}$. Univariate nonparametric statistics, *e.g.*, Kolmogorov-Smirnov statistic (Smirnov (1939)), rely on empirical estimates of the underlying distributions or related (pseudo)-metrics, see Appendix section B.2. The null hypothesis $\mathscr{H}_0$ is rejected if obtaining 'large' values of these statistics, *i.e.*, under 'large deviations' of the two random samples. For multivariate observations, natural empirical counterparts of their distributions are for instance

$$\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbf{X}_i} \text{ and } \widehat{\nu}_m = \frac{1}{m}\sum_{j=1}^{m} \delta_{\mathbf{Y}_j} , \tag{2.1.2}$$

where δ_x is the Dirac mass at any point x, or empirical versions of the *c.d.f.* , quantiles, copulas, depths, *etc.* Various classic (pseudo-)metrics measuring dissimilarity between two probability distributions are: chi-square distance, Kullback-Leibler divergence, Hellinger distance, Kolmogorov-Smirnov distance. Refer to Rachev (1991) for a comprehensive review. In minimax testing formulations, the alternative corresponds to the underlying distributions being different and separated in a metric sense, see *e.g.*, Lam-Weil et al. (2022) for local minimax separation rate defined by L_1-norm for discrete distributions, Carpentier et al. (2018) for L_2-norm in sparse linear regression. We refer in particular to Albert et al. (2021); Berrett et al. (2021) for independence testing and to Baraud (2002); Ingster and Suslina (2000); Lepski and Spokoiny (1999) for goodness-of-fit testing. See in particular Ingster and Suslina (2003) for a comprehensive overview regarding Gaussian models. The example below formulates a typical statistical test known as the location/shift.

Example 6. (LOCATION TEST IN $\mathbb{R}^d$) *In (semi)parametric testing, by considering P_1, $P_2 \in \mathscr{P}$ a probabilistic model, such that $G(t) = P_1(t-\theta_1)$, $H(t) = P_2(t-\theta_2)$, with parameters θ_1, $\theta_2 \in \mathbb{R}^d$, with $d \in \mathbb{N}^*$, the location problem is formulated as*

$$\mathscr{H}_0 : \ \theta_1 = \theta_2 \quad vs. \quad \mathscr{H}_1 : \ \theta_1 \neq \theta_2 \ .$$

The simplest form is usually presented when supposing P_1, P_2 known and equal. It recovers the Hotelling's T^2-test for the equality of means for Gaussian distributions.

While statistics can be constructed for a particular probabilistic model, *e.g.* Gaussian, Elliptical models, this manuscript focuses on nonparametric formulations for which obtaining statistical guarantees is possible. Precisely, we are interested in (asymptotic) consistency, (asymptotic) control of both statistical errors (type-I and type-II), independence of the statistics null distribution to the underlying model, independence of the test statistics to the transformations of the model under the alternative (also known as ancillary statistics Fisher (1925)), unbiasedness of the test statistic. Refer to Appendix section B.2 for details. Refer to classic books Gibbons and Chakraborti (2011); Lehmann and Romano (2005); Sheskin (2011); van der Vaart (1998) for comprehensive reviews of theory, methodologies and statistics in the field of (nonparametric) hypothesis testing.

2.1.2 Multivariate generalizations of classic statistics

The statistics gathered in this section are multivariate extensions of typical univariate statistics that are recalled in the Appendix section B.

A generalization of the Wald-Wolfowitz runs statistic: Friedman and Rafsky (1979). This article is one of the first that published a nonparametric generalization. It modeled the two samples through graphs structures that are constructed as follows. A weighted graph is formed by merging

the two samples, where the nodes represent the observations and the edges are weighted by the Euclidean distance, or by a generalized dissimilarity, between the two related points. The statistic relies on considering the *Minimal Spanning Tree* (MST), that can be formulated as a subgraph passing through all the points while minimizing the total weight, without any cycle. It is unique if there are no ties. The Wald-Wolfowitz runs (WW) statistic R is hence generalized by considering the total number of subtrees obtained when constructing the two-sample MST and removing all the edges linking two observations from different samples, see Eq. (B.2.5) for the univariate definition in the Appendix. The null $\mathscr{H}_0$ is rejected for a small number of obtained runs. In fact, R can be interpreted as a correlation coefficient between the interpoints distances and the sample identities, allowing to prove its asymptotic normality. Additionally, the null distribution is shown to be independent on the sample's distribution if it is conditioned by the *r.v.* corresponding to the number of edge pairs having common nodes. While performing poorly for small dimensions on (log)normal location and scale distributions, it is shown to have a competitive power for the location when the dimension of the data increases. However, the obtained statistic relies on the coordinate representations of the points and therefore depends on their scaling choices. Hence, the distribution of the statistic under the alternative is biased relatively to these changes, leading to various power performances.

A generalization of the Kolmogorov-Smirnov statistic: Præstgaard (1995). The Kolmogorov-Smirnov (KS) statistic is valued on the univariate mapped observations by means of a scoring functions f, see Appendix (B.2.4) for the formal definition. Consider a sequence of classes $\mathscr{F}_N \subset \mathscr{F}$, of measurable functions mapping $\mathscr{X}$ to $\mathbb{R}$. Suppose $\mathscr{F}_N$ 'converges' to $\mathscr{F}_0 \subset \mathscr{F}$ as $N \to \infty$. The generalized statistic indexed by $\mathscr{F}_N$ is given by

$$D_{n,m} = \sup_{f \in \mathscr{F}_N} \left| \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{X}_i) - \frac{1}{m} \sum_{j=1}^{m} f(\mathbf{Y}_j) \right| . \tag{2.1.3}$$

The author proved the asymptotic consistency of both the permutation and bootstrap tests under fixed and local alternatives as soon as: (i) assumptions on $\mathscr{F}$ are satisfied such that convergence theorems for the obtained empirical processes hold (see Section 1 and van der Vaart (1998)), (ii) the sup norm over $\mathscr{F}_0$ of $(1/n)\sum_{i\leq n} f(\mathbf{X}_i) - (1/m)\sum_{j\leq m} f(\mathbf{Y}_j)$ is not null.

A smooth generalization of the Neyman test: Zhou et al. (2017). Assume the marginals of both random variables $\mathbf{X}$ and $\mathbf{Y}$ are independent, the proposed statistic learns the optimal linear direction to project the variables. This statistic is based on the KS statistic. Let $\Psi = (\psi_k)_{k\leq d}$ a d-variate orthonormal function such that the coordinates satisfy $\int_0^1 \psi_k(x)\psi_\ell(x)dx = \delta_{k,\ell}$, for all $k, \ell \leq d$, then

$$T_{n,m}(d') = \sqrt{\frac{nm}{n+m}} \sup_{a \in \mathscr{S}^{d-1}} \widehat{\Psi}_a(d') , \tag{2.1.4}$$

where $\widehat{\Psi}_a(d') = \max_{k\leq d'} |(1/m)\sum_{j\leq m} \psi_k(\widehat{G}_n^a(a^T Y_j))|$ with $\widehat{G}_n^a$ is the empirical distribution of the univariate sample $\{a^T\mathbf{X}_1, \ldots, a^T\mathbf{X}_n\}$, for all $a \in \mathscr{S}^{d-1}$ *i.e.* in the unit sphere, and where $d' \leq d$ allows possible truncation. A bootstrap procedure is derived, weighting the d' coordinates by a *i.i.d.* standard Gaussian sequence that is independent on the two samples. For the smoothed counterpart of the statistic, the asymptotic consistency is guaranteed. Numerical results show high power in detecting local features or high-frequency components in comparison with Baringhaus and Franz (2004). However, these asymptotic results rely on strong assumptions on the probabilistic model.

A L_1-based test statistic: Biau and Gyorfi (2005). The authors introduce the statistic based on the L_1-distance between the two empirical distributions restricted to a partition of the feature space,

in the case of balanced samples (*i.e.* $n = m$). Consider a finite partition $\mathscr{P}_n = \{A_{n1}, \ldots, A_{np_n}\}$ of $\mathbb{R}^d$ supposed to increase with n, with $p_n \in \mathbb{N}^*$. The statistic is defined as follows

$$T_n = \sum_{i=1}^{p_n} |\widehat{\mu}_n(A_{ni}) - \widehat{\nu}_n(A_{ni})| \,, \tag{2.1.5}$$

where the empirical measures are defined according to the empirical maps (2.1.2). Under the alternative, the authors provide large deviation bounds *à la* Chernoff, independent on the underlying distributions, and prove strong consistency[1]. The asymptotic null distribution is not distribution-free and shows a smaller critical value for the test to be asymptotically consistent.

A multivariate extension of the energy statistic: Szekely and Rizzo (2004). The authors propose a straightforward extension of the univariate energy statistic by means of the Euclidean distance in $\mathscr{Z} \subset \mathbb{R}^d$,

$$\mathscr{E}_{n,m} = \frac{mn}{m+n}\left(\frac{2}{nm}\sum_{i,j\leq n,m}\|\mathbf{X}_i - \mathbf{Y}_j\| - \frac{1}{n^2}\sum_{i,j\leq n}\|\mathbf{X}_i - \mathbf{X}_j\| - \frac{1}{m^2}\sum_{i,j\leq m}\|\mathbf{Y}_i - \mathbf{Y}_j\|\right) \,, \tag{2.1.6}$$

where $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^d$. If the *r.v.* $\mathbf{X}$ and $\mathbf{Y}$ have finite variance, the statistic (2.1.6) is consistent in power against fixed alternatives. The asymptotic null distribution is also derived and depends on the kernel of the statistic. This contribution is accompanied with a package named `energy`, implemented in the statistical software environment R.

A modified energy test statistic based on projections: Baringhaus and Franz (2004). The statistic introduced is based on the L_2-distance between the distribution functions, once the multivariate variables are projected on the unit sphere. Its empirical counterpart is defined by

$$T_{n,m} = \gamma_d \frac{mn}{m+n}\int_{\mathscr{S}^{d-1}}\int_{\mathbb{R}}(\widehat{G}_n^a(t) - \widehat{H}_m^a(t))^2 dt d\lambda(a) \,, \tag{2.1.7}$$

for all $a \in \mathscr{S}^{d-1}$, where $\widehat{G}_n^a$ and $\widehat{H}_m^a$ are the empirical distributions of the samples $\{a^T\mathbf{X}_1, \ldots, a^T\mathbf{X}_n\}$ and $\{a^T\mathbf{Y}_1, \ldots, a^T\mathbf{Y}_m\}$, and $\gamma_d = \sqrt{\pi}(d-1)\Gamma((d-1)/2)/2\Gamma(d/2)$ the normalizing constant such that: $\|x\| = \gamma_d \int_{\mathscr{S}^{d-1}} |a^T x| dx d\lambda(a)$ with λ the uniform measure on the sphere. The asymptotic null distribution is obtained in Th. 2.2. It is the continuous equivalent to the energy statistic, see Szekely (2003). The statistic is consistent and invariant to orthogonal linear transformations. For the location problem, numerical experiments show high power for the Gaussian setting but sensitivity to general distributions, as well as low power for detecting local features or high-frequency components. This contribution is accompanied with a package named `cramer` is implemented in the statistical software environment R.

A generalization of Mann-Whitney-Wilcoxon: Clémençon et al. (2009). The R-statistic is generalized by a bipartite ranking approach, for which a *scoring function* s is learnt as to map the multivariate observations to $\mathbb{R}$. By considering a class $\mathscr{S} = \{s : \mathscr{Z} \to \mathbb{R} \cup \{+\infty\},\ s \text{ measurable}\}$, MWW is generalized thanks to

$$\widehat{W}_{n,m}(s) = \sum_{i=1}^{n} \text{Rank}(s(\mathbf{X}_i)) \,, \tag{2.1.8}$$

[1] A strong consistent test is consistent for all the points in the null hypothesis for the acceptance and in the alternative for the rejection, with probability one.

where $\text{Rank}(t) = \sum_{i=1}^{n} \mathbb{I}\{s(\mathbf{X}_i) \leq t\} + \sum_{j=1}^{m} \mathbb{I}\{s(\mathbf{Y}_j) \leq t\}$ for all $t \in \mathbb{R}$. This approach is motivated by the univariate relation of MWW to the *Area Under the* ROC *Curve* (AUC). Hence, once the optimal scoring function is learnt on a subsample of the whole dataset *s.t.* it maximizes the AUC, it is then used to rank the remaining samples of observations in order to compute the univariate MWW hypothesis test. Asymptotic consistency and Gaussian null distribution of the test statistic are proved (Th. 2). The procedure is accompanied by promising numerical results showing high power.

2.1.3 Statistics based on kernel methods

Both statistics in the sequel are based on dissimilarity measures in Hilbert space embeddings, mapping a probability distribution into a *Reproducing Kernel Hilbert Space* (RKHS). Typical kernel methods are applied on the mapped distributions. Briefly, this method extends the Euclidean inner product in $\mathscr{Z}$ to high-dimensional spaces by means of a kernel function $k(\cdot,\cdot)$[2] related to the mapping $P \mapsto \mu_P = \int_{\mathscr{Z}} k(z,\cdot) dP(z)$, where P is a distribution. In particular, it allows to represent the expectation as an inner product $\mathbb{E}_{Z\sim P}[f(Z)] = \langle f, \mu_P \rangle$, by the reproducing property of the space, see *e.g.* Berlinet and Thomas-Agnan (2004); Schölkopf and Smola (2002); Schölkopf et al. (2003) for further references. In the following, we denote by $(\mathscr{F}, \langle\cdot,\cdot\rangle_{\mathscr{F}})$ a RKHS.

Maximum Mean Discrepancy statistic (MMD): Gretton et al. (2007, 2012a). The MMD two-sample statistic is defined as the uniform bound in expectations over functions in the unit ball of an RKHS space. Denote it by $\mathscr{F}$ that is associated to the Mercer kernel $k(\cdot,\cdot)$. This classic test has gained popularity thanks to its simple definition, its possible implementation for various data structures, as well as its adaptability to other two-sample statistical tests. The functional definition is

$$\text{MMD}(G,H) = \sup_{f\in\mathscr{F}} |\mathbb{E}[f(\mathbf{X})] - \mathbb{E}[f(\mathbf{Y})]| \ , \tag{2.1.9}$$

where the kernels are supposed to be *characteristic*, *e.g.* Gaussian, Laplace, such that the equivalence holds: $\text{MMD}(G,H) = 0$ *iff* $G = H$, see Theorem 5. From this definition, both a biased and an unbiased empirical counterparts of the square statistic have been studied. The squared counterpart can be viewed as a L_2 distance-based statistic between kernel density estimators. The results are mainly on consistency and large deviation bounds with fixed kernels where the constants in the bounds are independent of the underlying distributions. The third test is based on the asymptotic distribution of the unbiased estimate. Although the bounds are distribution-free, the null distribution is not and its estimation requires data-driven procedures, here for instance thanks to the Pearson curves (Section 5) *i.e.* estimating the moments of the statistic, see Johnson et al. (1994) Section 18.8. Additionally, empirical results show the sensitivity of the statistics *w.r.t.* the kernel's bandwidth especially for small sample sizes, requiring data-splitting in order to optimize it and then to perform the testing procedure. Alternatively, the heuristic choice is the median inter-sample distance, but empirically shows low power while not being theoretically proved, see Gretton et al. (2012a). Lastly, in order to achieve low algorithmic complexity, the authors propose a *linear time* test ($\mathscr{O}(n+m)$) thanks to the unbiased version. The central limit theorem applied to the linear and to the unbiased statistics, proves Gaussian asymptotic laws, where the variances depend on the unknown distributions G and H (Cor. 16). Tail bounds using Hoeffding (1963) depend on the class of bounded kernels (*e.g.* Th. 7 biased statistic, Th. 10 unbiased statistic, Th. 15 linear statistic). Hence five test statistics are available for implementation: the biased (Cor. 9) and four unbiased counterparts with sequentially the thresholds computed *via* Hoeffing's deviation bound (Cor. 11), bootstrap sampling, moment matching Pearson

[2] If k supposed positive definite, there exists a unique map (not necessarily known) $\psi : \mathscr{Z} \to \mathscr{F}$, such that $k(x,y) = \langle \psi(x), \psi(y) \rangle$, see Aronszajn (1950).

curves (Section 5) and lastly the linear time statistic (Cor. 16). We refer to Chwialkowski et al. (2016); Gretton et al. (2009, 2012b); Li et al. (2017); Schrab et al. (2021) for extensions of the MMD-based statistic.

Maximum Kernel Fisher ratio statistic: Bach et al. (2008). Following the MMD, the proposed statistic is based on L_2-distance between kernel estimators, weighted by the covariance structure of the samples. Precisely, the estimator of the operators: $\widehat{\Sigma}_W$ the pooled/within-class covariance and $\widehat{\Sigma}_B$ the between-class covariance, are based on the covariance of each sample Σ_X, Σ_Y. The statistic is the normalized version of

$$T_{n,m} = N \max_{f \in \mathscr{F}} \frac{\langle f, \widehat{\Sigma}_B f \rangle_{\mathscr{F}}}{\langle f, (\widehat{\Sigma}_W + \gamma_N \mathbb{I}_d) f \rangle_{\mathscr{F}}} , \tag{2.1.10}$$

with $N = n+m$, where $\mathbb{I}_d$ the identity matrix and $\{\gamma_i\}_{i \leq N}$ a sequence in $\mathbb{R}_+^*$, typically converging to 0. The 'normalization' of $T_{n,m}$ aims to enhance the power: for instance setting $\gamma_n = 0$ and choosing linear kernel yields to the T^2-Hotelling statistic, see Lehmann and Romano (2005). Classic asymptotic results are proved for two modelings of the sequence γ_N: (i) constant $\gamma_N = \gamma$ and (ii) depending on Σ_X, Σ_Y. The limit distribution under $\mathscr{H}_0$ is proved to be Gaussian: for (i) the parameters depend on Σ_W and on γ (Theorem 1), for (ii) it is standard normal (Theorem 3). The consistency in power is proved for fixed alternative as well as for local ones[3] to speed the convergence rate, for both (i, ii) (Th. 5). Here the sequence of alternatives are defined by the convergence to 0 of the χ^2-divergence between the probability laws (Prop. 4). Notice that although the results under (ii) are appealing, conditions are necessary *w.r.t.* the decay rate of γ_N, that depends on the underlying distributions of the samples.

Remark 1. (ENERGY DISTANCE AND KERNELS) *As pointed out Ramdas et al. (2015), energy distances and kernels are closely related by a simple relation $D(x,y) = (k(x,x) + k(y,y))/2 - k(x,y)$, for a distance D and a kernel k, see the comprehensive study in Sejdinovic et al. (2013).*

2.1.4 Statistics based on optimal transport distances

Methods introduced in hypothesis testing using optimal transport theory are based on the comparison of probability measures in metric spaces *via* transport measures, such as the family of Wasserstein distances, see Villani (2009). In particular, the p-Wasserstein distance at power $p \geq 1$, also defined as the Mallow's distance in the statistical literature, is the optimum of the transportation problem, where it has a linear objective *w.r.t.* a polyhedral feasible set. These tests can be seen as generalizations of kernel and energy based tests, see for instance Feydy et al. (2018); Ramdas et al. (2015). Although they rely on a similar concept, this extension allows for broader analysis and interpretation as the Wasserstein distance is considered instead of the Euclidean one.

Smooth Wasserstein statistic: Ramdas et al. (2015). Consider $\mathscr{Z} = \mathbb{R}^d$ and suppose the distributions G and H have finite p-moments, for $p \in [1, \infty)$. The authors introduced a statistic based on the Wasserstein distance with an additional entropy penalty/regularization as the solution of the following problem

$$T_\lambda = \underset{T \in U_{n,m}}{\arg\min}\ \lambda \langle T, M_{XY} \rangle - E(T) , \tag{2.1.11}$$

[3] See Lehmann and Romano (2005). Local alternatives are defined as a sequence of alternatives tending to the null as $N \to \infty$. The rate is such that the limit *r.v.* is nondegenerate.

where $\lambda > 0$ and $U_{n,m}$ is defined as the polytope of nonnegative matrices *s.t.* their row (*resp.* columns) equal to $\mathbf{1}_n/n$ (*resp.* $\mathbf{1}_m/m$), $\mathbf{1}_n$ being the unit n-dimensional vector. The pairwise distances of $\mathbf{X}$ and $\mathbf{Y}$ at power p is $M_{XY} = (\|\mathbf{X}_i - \mathbf{Y}_j\|^p)_{i,j\leq n,m}$. The entropy of T is defined as a discrete joint probability distribution: $E(T) = -\sum_{i,j\leq n,m} T_{ij}\log(T_{ij})$. In fact, the p-Wasserstein distance between the empirical distributions is $W_p(\widehat{G}_n, \widehat{H}_m) = \min_{T\in U_{n,m}}\langle T, M_{XY}\rangle$. This formulation is strongly convex, hence admits a unique solution, such that the optimal statistic is dependent on λ and is a diagonal scaling of $e^{-M_{XY}}$. This objective function balances both the energy-based and transport-based statistics, by tailoring the weight of the penalty. The authors do not study classic properties related to the test statistic.

Rank-based statistic with measure transportation: Deb and Sen (2019). The class of statistics introduced relies on an extension of the ordered rank variables thanks to optimal transport, resulting to a *population multivariate rank map*. Suppose the measures μ, ν are absolutely continuous *w.r.t.* the Lebesgue measure in $\mathbb{R}^d$, and consider a pre-specified reference measure Λ. The *rank map* is the unique solution of the Monge transportation problem, where this map is the push-forward of the mixture measure of the two samples to the reference measure Λ. It therefore depends on the studied hypothesis. As both underlying distributions are unknown, the empirical rank map $\widehat{R}_{n,m}$ is the optimal mapping from the empirical mixture measures (see (2.1.2)) to the chosen Λ. This multivariate rank map allows for various choices of Λ, such as Gaussian, Uniform in $[0,1]^d$, see Section D.2 therein for other examples. For instance, it is chosen as the d-dimensional Halton sequence of same size of the sample. The proposed method relies on using the optimal map as a *plug-in* variable for multivariate tests such as Szekely (2003); Székely et al. (2007). In particular, the rank energy statistic of Eq. (2.1.6) is defined as

$$\mathscr{RE}_{n,m} = \frac{mn}{m+n}\Big(\frac{2}{nm}\sum_{i,j\leq n,m}\|\widehat{R}_{n,m}(\mathbf{X}_i) - \widehat{R}_{n,m}(\mathbf{Y}_j)\| - \frac{1}{n^2}\sum_{i,j\leq n}\|\widehat{R}_{n,m}(\mathbf{X}_i) - \widehat{R}_{n,m}(\mathbf{X}_j)\| - \frac{1}{m^2}\sum_{i,j\leq m}\|\widehat{R}_{n,m}(\mathbf{Y}_i) - \widehat{R}_{n,m}(\mathbf{Y}_j)\|\Big)\,. \quad (2.1.12)$$

This method yields to exact distribution-free tests *i.e.* for all sample sizes (Lemma 4.3), explicit asymptotic distributions under the alternative (Theorem 4.3) and are consistent to fixed alternatives (Theorem 4.4), as it inherits of the attractive properties of rank statistics. Also, the optimization of the rank map can be linearly formulated, resulting in a computationally feasible procedure. This method of multivariate rank map was also extended to the T^2-Hotelling test in Deb et al. (2021).

2.2 Bipartite ranking

Bipartite ranking has gained popularity in the Machine Learning community thanks to its adaptability to numerous application fields, ranging for instance from information retrieval, recommendation systems, to biomedical studies. As a particular case of *ranking* problems, the goal is to learn an optimal *scoring function*, such that univariate mapping of the observations obtained by this function induces an ordering minimizing a statistical loss. Precisely, the observations are considered to have a binary label, referred to as either 'positive' or 'negative', yielding a comparison of each instance from one class to another. The resulting pairwise loss corresponds to the probability of misranking pairs of observations by a scoring function, drawn at random from each label. In fact, it is intimately related to *Receiver Operating Characteristic* (ROC) analysis as it provides a quality measure on the class of scoring functions. The bipartite ranking risk is known to be equal to one minus the *Area Under the* (ROC) *Curve* (AUC). This is the direction we chose for the chapter and more generally

for the works gathered in this manuscript. We refer for extensions to *multipartite ranking* (*i.e.* when the number of different labels under study is larger than 3) to the works of Clémençon and Robbiano (2015); Clémençon et al. (2013b). In this section, the probabilistic framework will be presented with its intrinsic relation to the ROC analysis. Optimization formulations are next detailed and finally the state-of-the-art algorithms are reviewed.

2.2.1 Probabilistic formulation

In its generic formulation, consider the input variable $\mathbf{Z}$ defined on the probability space $(\Omega, \mathscr{A}, \mathbb{P})$ and valued in the input/feature space $\mathscr{Z}$, associated to its binary label ζ valued in $\{-1, +1\}$. A common and fundamental choice for $\mathscr{Z}$ is a subset of the Euclidean space $\mathbb{R}^d$, with $d \geq 2$. Reformulating the heuristic of *ranking* leads to the comparison of the variables $\mathbf{Z}$ and $\mathbf{Z}'$ of *resp.* labels $\zeta = 1$ and $\zeta' = -1$. Notice that by considering the posterior probability $\eta(z) = \mathbb{P}\{\zeta = 1 \mid \mathbf{Z} = z\}$, the problem is completely defined by the couple (F, η), where $F(dz)$ is the marginal of the *r.v.* $\mathbf{Z}$. We suppose that the probability of the positive label is $p = \mathbb{P}\{\zeta = 1\} \in (0,1)$. With this setup at hand, the goal is to learn the optimal *scoring function* s from a class of candidates $\mathscr{S} = \{s : \mathscr{Z} \to \mathbb{R} \cup \{+\infty\},\ s \text{ measurable}\}$, such that it minimizes the bipartite ranking risk defined as follows.

Definition 7. *The bipartite ranking risk/error of a scoring function $s(z)$ w.r.t. the distributions of $s(\mathbf{Z}') \mid \{\zeta = -1\}$ and $s(\mathbf{Z}) \mid \{\zeta = 1\}$, associated to the $0-1$/binary loss, is defined by:*

$$L(s) = \mathbb{E}[\mathbb{I}\{s(\mathbf{Z}') > s(\mathbf{Z})\} \mid \zeta' = -1,\ \zeta = 1] + \frac{1}{2}\mathbb{P}\{s(\mathbf{Z}') = s(\mathbf{Z}) \mid \zeta' = -1,\ \zeta = 1\} \tag{2.2.1}$$

where the tights are broken at random. The optimal scoring function minimizing the risk s^* is defined such that $L(s^*) = \inf_{\mathscr{S}} L =: L^*$.

This definition highlights the importance of the induced order obtained by the image of the observations by a scoring function, instead of the value of the score itself. The ideal scoring aims to attribute the higher ranks to the positive labels with high probability. The following Proposition states that the posterior probability η achives the minimum bipartite ranking error, see Clémençon et al. (2008).

Proposition 8 (Example 1, Clémençon et al. (2008))**.** *Consider $s \in \mathscr{S}$ and $\eta(z) = \mathbb{P}\{\zeta = 1 \mid \mathbf{Z} = z\}$. The bipartite ranking excess of risk of the scoring function s equals to:*

$$L(s) - L^* = \mathbb{E}[|\eta(\mathbf{Z}') - \eta(\mathbf{Z})|\mathbb{I}\{(s(\mathbf{Z}) - s(\mathbf{Z}'))(\eta(\mathbf{Z}) - \eta(\mathbf{Z}')) < 0\}]\,, \tag{2.2.2}$$

where the Bayes risk is defined by $L^* = \mathbb{E}[\min(\eta(\mathbf{Z}'), \eta(\mathbf{Z}))] - \mathbb{E}[\eta(\mathbf{Z})]^2$ and the r.v. $\mathbf{Z}'$ and $\mathbf{Z}$ are i.i.d. drawn from F.

In the following paragraph, the statistical risk as defined in (2.2.1) is shown to be intimately related to ROC analysis as it equals to one minus the corresponding AUC. Precisely, the goal is to highlight how to measure the *quality* of a scoring function thanks to the obtained ROC criterion. This approach has become a traditional modeling for the study of bipartite ranking problems, leading to maximizing empirical versions of the AUC criterion, see *e.g.* Agarwal et al. (2005) or Clémençon et al. (2008). We refer to the Appendix section B.3 for the univariate introduction.

Bipartite ranking optimization and optimal elements *via* the ROC analysis. The scoring function can be seen as a natural way of defining a total preorder[4] on $\mathscr{Z}$ by mapping it with the natural

[4] A preorder $\preccurlyeq$ on a set $\mathscr{Z}$ is a reflexive and transitive binary relation on $\mathscr{Z}$. It is said to be *total*, when either $z \preccurlyeq z'$ or else $z' \preccurlyeq z$ holds true, for all $(z, z') \in \mathscr{Z}^2$.

order on $\mathbb{R} \cup \{+\infty\}$. The ability of a given $s(z)$ to discriminate between a 'positive' and a 'negative' variables, can be evaluated through the Probability-Probability (P-P) plot defined as the functional criterion depending on s, namely the ROC curve.

Definition 9 (Definition 4, Clémençon and Vayatis (2009b))**.** *The Receiver Operating Characteristic (*ROC*) curve of a scoring function $s(z)$ is the parametric curve:*

$$\mathrm{ROC_s} : \mathrm{t} \in \mathbb{R} \mapsto (1 - \mathrm{H_s(t)}, 1 - \mathrm{G_s(t)}) \ ,$$

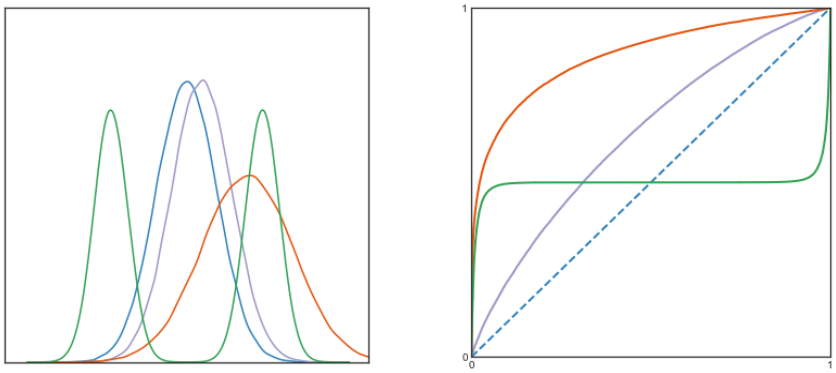
valued in $[0,1]^2$, where for all $t \in \mathbb{R}$,

$$\begin{aligned} G_s(t) &= \mathbb{P}\{s(\mathbf{Z}) \leq t \mid \zeta = 1\} \\ H_s(t) &= \mathbb{P}\{s(\mathbf{Z}) \leq t \mid \zeta = -1\} \ . \end{aligned}$$

The function $t \mapsto 1 - G_s(t)$ is defined as the True Positive Rate (TPR) while $t \mapsto 1 - H_s(t)$ is defined as the False Positive Rate (FPR). Moreover, by the univariate definition Eq. (B.3.2) (see Appendix section B), the ROC curve can be parametrized as follows

$$\mathrm{ROC(s,\cdot)} : \alpha \in (0,1) \mapsto 1 - \mathrm{G_s} \circ \mathrm{H_s^{-1}}(1-\alpha) =: \mathrm{ROC_{H_s,G_s}}(\alpha) \ . \tag{2.2.3}$$

As for the univariate definition, we consider the possible jumps of the curve connected by line segments at the degenerate points of the functions G_s and H_s, such that the considered ROC curve is always continuous, see Clémençon and Vayatis (2009b). Inherited by the definition of the bipartite ranking risk (Eq. (2.2.1)), a *good* scoring rule $s(z)$ can be interpreted as such that the distribution G_s is 'as stochastically larger as possible'[5] than H_s. It is graphically understood as the more the $\mathrm{ROC(s,\cdot)}$ curve is close to the point of coordinates $(0,1)$, the better the scorer $s(z)$. Figure 2.1 illustrates how the concept of ROC curve offers a visual tool to examine the differences between two distributions in a pivotal manner.



a. Probability distributions b. ROC curves

Figure 2.1. Examples of pairs of distributions and their related ROC curves. The distribution H is represented in blue and three examples of G distributions are in purple, orange and green, like the associated ROC curves.

[5]Given two distribution functions $H(dt)$ and $G(dt)$ on $\mathbb{R} \cup \{+\infty\}$, it is said that $G(dt)$ is *stochastically larger* than $H(dt)$ *iff* for any $t \in \mathbb{R}$, we have $G(t) \leq H(t)$. We then write: $H \leq_{sto} G$. Classically, a necessary and sufficient condition for G to be stochastically larger than H is the existence of a coupling $(\mathbf{X}, \mathbf{Y})$ of (G,H), *i.e.* a pair of random variables defined on the same probability space with first and second marginals equal to H and G respectively, such that $\mathbf{X} \leq \mathbf{Y}$ with probability one.

Definition 10. *A scoring function is optimal in the sense of the minimization of the bipartite risk error, denoted by* $s^* : \mathscr{Z} \mapsto \mathbb{R} \cup \{+\infty\}$*, if it induces the same order as the function* $z \mapsto \eta(z)$*, i.e. :*

$$\forall (z,z') \in \mathscr{Z}^2 \quad (\eta(z) < \eta(z') \quad or \quad \Psi(z) < \Psi(z')) \implies s^*(z) < s^*(z') \,, \tag{2.2.4}$$

where $\Psi(z) = (p/(1-p))\eta(z)/(1-\eta(z))$ *is the likelihood ratio.*

When interpreting the ROC curve as the power graph of the hypothesis test: $\mathscr{H}_0 : \zeta = 1$ *v.s.* $\mathscr{H}_1 : \zeta = -1$ based on $\mathbf{Z}$, the Neyman-Pearson Lemma (*e.g.* Lehmann and Romano (2005)) states that likelihood ratio test $\Psi(\mathbf{Z})$ is the uniformly most powerful test among all the tests based on $\mathbf{Z}$. The characterization of the set of optimal elements is therefore defined as follows.

Proposition 11 (Proposition 2, Clémençon and Vayatis (2008))**.** *The class of optimal scoring functions is defined as the set:*

$$\mathscr{S}^* = \left\{ s \in \mathscr{S} \;\; s.t. \; for \; all \; z, \, z' \; in \; \mathscr{Z} : \;\; \eta(z) < \eta(z') \Rightarrow s^*(z) < s^*(z') \right\} \,. \tag{2.2.5}$$

Consequently, for an optimal scoring function, there exists a strictly increasing map defined by $T : Im(s) \to \mathbb{R}$, such that $s^* = T \circ \eta$, where $Im(s) \subset \mathbb{R}$ is the image set induced by s. A generic form of the optimal bounded scoring function is obtained in the same article.

Proposition 12 (Proposition 3, Clémençon and Vayatis (2008))**.** *A bounded scoring function* s^* *is optimal iff there exists a nonnegative integrable function* w *and a continous r.v.* V *valued in* $(0,1)$ *such that, for all* $t \in \mathscr{Z}$*:*

$$s^*(t) = \inf_{\mathscr{Z}} s^* + \mathbb{E}[w(V)\mathbb{I}\{\eta(t) > V\}] \,. \tag{2.2.6}$$

The function w is related to the scale of the scoring function *i.e.* to $Im(s)$, while V to the inverse of the function T. With these optimality characterizations at hand, and by recalling that the ROC curve is invariant by any strictly increasing transform, we have that the ROC curve is invariant for all optimal scoring functions in $\mathscr{S}^*$. Hence, for all $s^* \in \mathscr{S}^*$, the optimal ROC curve is equal to $\mathrm{ROC}(s^*, \cdot) = \mathrm{ROC}(\eta, \cdot) =: \mathrm{ROC}^*(\cdot)$. It is non-decreasing, concave and hence always above the main diagonal of the unit square. Importantly, the ROC curve induces a partial preorder on the set of all scoring functions. For all pairs (s_1, s_2), one can define s_2 as more accurate than s_1 if $\mathrm{ROC}(s_1, \alpha) \leq \mathrm{ROC}(s_2, \alpha)$ for all $\alpha \in [0,1]$. We refer to Clémençon and Vayatis (2009b) for additional basic properties of ROC curves. The sup norm is usually used to measure the distance between two ROC curves and in particular, of the deviation for a given scoring function $s(z)$ *w.r.t.* the optimal one.

$$d_\infty(s, s^*) = \sup_{\alpha \in (0,1)} |\mathrm{ROC}(s, \alpha) - \mathrm{ROC}^*(\alpha)| \,. \tag{2.2.7}$$

We highlight that this distance is measured in the ROC space and not directly between the scoring functions. However in practice, ROC^* is unknown and there is no statistical counterpart of the functional loss (2.2.7). In Clémençon and Vayatis (2009b, 2010), bipartite ranking was shown similar to a superposition of cost-sensitive classification problems being 'discretized' through adaptive steps. This allows for applying empirical risk minimization with statistical guarantees in the d_∞-sense, at the price of an additional bias term inherent to the approximation step. Alternatively, the performance of a scoring rule s can be measured by means of the L_1-norm in the ROC space. Observing that, in this case, the loss can be decomposed as follows

$$d_1(s, s^*) = \int_0^1 |\mathrm{ROC}(s, \alpha) - \mathrm{ROC}^*(\alpha)| \mathrm{d}\alpha = \int_0^1 \mathrm{ROC}^*(\alpha) \mathrm{d}\alpha - \int_0^1 \mathrm{ROC}(s, \alpha) \mathrm{d}\alpha \,, \tag{2.2.8}$$

minimizing the L_1-distance to the optimal ROC curve boils down to maximizing the area under the curve $\mathrm{ROC}(s, \cdot)$, defined below.

Definition 13. *Let $s(z)$ be a scoring function, the Area Under the* ROC *Curve (*AUC*) is defined as:*

$$\mathrm{AUC}(\mathrm{s}) = \int_0^1 \mathrm{ROC}_{\mathrm{H_s,G_s}}(\alpha)\mathrm{d}\alpha =: \mathrm{AUC}_{\mathrm{H_s,G_s}} \ . \tag{2.2.9}$$

We denote $\mathrm{AUC}^* = \mathrm{AUC}(\mathrm{s}^*)$*, with* $s^* \in \mathscr{S}^*$.

In fact, it completely recovers the equation (2.2.1) as follows

$$\begin{aligned} \mathrm{AUC}(\mathrm{s}) &= \mathbb{P}\{s(\mathbf{Z}') < s(\mathbf{Z}) \mid \zeta = 1, \zeta' = -1\} + \frac{1}{2}\mathbb{P}\{s(\mathbf{Z}') = s(\mathbf{Z}) \mid \zeta = 1, \zeta' = -1\} & (2.2.10)\\ &= 1 - L(s) \ . & (2.2.11) \end{aligned}$$

The scalar performance criterion AUC(s) defines a total preorder on $\mathscr{S}$ and its maximal value is attained on the set $\mathscr{S}^*$ of optimal value AUC^*. Hence, when considering the binary loss, the functional ROC curve, or the AUC as a scalar summary, are goldstandard measures for the quality of a scoring function $s(z)$ in the context of bipartite ranking.

A statistical formulation. Based on independent random copies $\{(\mathbf{Z}_i, \zeta_i)_{i\leq N}\}$, with $N \in \mathbb{N}^*$, the goal of bipartite ranking is to learn how to *score* any new sample $\mathbf{Z}_{N+1}, \ldots, \mathbf{Z}_{N+k}$, such that it minimizes the empirical counterpart of the expected loss function $L(s)$, or equivalently to maximize the empirical AUC(s). Notice that the new instances can be either 'positive' or 'negative' when assuming no prior knowledge. Indeed, as in practice the distribution of the pair $(\mathbf{Z}, \zeta)$ is unknown, one has only access to random *i.i.d.* copies. The empirical counterpart of the AUC (equivalently of the bipartite ranking risk) for a scoring function s is given by

$$\begin{aligned} \widehat{\mathrm{AUC}}(s) &= \frac{1}{nm} \sum_{\{i,\, \zeta_i=+1\}} \sum_{\{j,\, \zeta_j=-1\}} \left(\mathbb{I}\{s(\mathbf{Z}_j) < s(\mathbf{Z}_i)\} + \frac{1}{2}\mathbb{I}\{s(\mathbf{Z}_j) = s(\mathbf{Z}_i)\} \right) & (2.2.12)\\ &= 1 - \widehat{L}(s) \ , & (2.2.13) \end{aligned}$$

where $n = \sum_{i\leq N} \mathbb{I}\{\zeta_i = +1\}$ and $m = \sum_{i\leq N} \mathbb{I}\{\zeta_i = -1\}$.

Hence, assumptions are necessary to ensure that the estimator converges to the expected criterion and, in particular, the subject of study is to be able to control the (uniform) fluctuations of $\widehat{L}(s) - L(s)$, when indexed by $\mathscr{S}$. Incidentally, this empirical formulation reveals the difficulty of such modeling: the pairwise comparison induces non-*i.i.d.* sums that do not allow for traditional optimization tools, *e.g.*, of empirical risk minimization, and takes the form of higher-order statistics such as U-statistics. Chapter 3 introduces these *unbiased* statistics. We define U-processes as collections of such statistics when indexed by the class of scoring functions $\mathscr{S}$. In the following section, computational approaches for optimizing the objective counterpart (and related) are detailed, particularly for leveraging pairwise empirical losses.

2.2.2 Optimization and state-of-the-art algorithms

This section briefly introduces extensions of the bipartite ranking risk by choosing the loss function to find an optimal global scoring element. The associated main algorithms are also presented, particularly those considered state-of-the-art for ranking problems. Broadly, three main approaches are proposed for learning the global optimal scoring function, either through (i) *plug-in* procedures, (ii) the minimization of the empirical risk or related estimators, or by (iii) transferring solutions of univariate binary classification (BC) or class probability estimation (CPE) models, to bipartite ranking.

Plug-in methods. The optimal result recalled in Proposition 11 motivates the direct statistical estimation of the likelihood ratio/posterior probability. This estimation is used as a scoring rule, see Devore and Lorentz (1993). For instance, plug-in rules based on partitioning the feature space are derived in Clémencon and Vayatis (2009), were defined as piecewise constant functions that maximize the AUC on each cell in the L_1-sense. This direction, however, quickly finds limitations in general contexts. It often requires a parametrization of the model and could provide estimators that are not consistent in the L_∞-sense. More generally, such approaches can suffer from the curse of dimensionality due to the possible complex structure of $\mathscr{X}$, see *e.g.* Devroye et al. (1996) Section 28.4. The convergence rate has been shown to depend on (and to decrease with) the dimension d, if $\mathscr{X} \subset \mathbb{R}^d$. Hence, possible approaches require sparsity assumptions or stronger assumptions like the independence of the marginals of the conditional distributions, see *e.g.* Guedj and Robbiano (2018); Li et al. (2013).

Minimization of the risk: How to bridge the gap to pairwise classification losses. Most of the bipartite ranking algorithms model the learning task through pairwise classification. By considering all the combinations of pairs $(\mathbf{Z}, \mathbf{Z}')$, these algorithms are of quadratic complexity. Formally, we recall the formulation studied in Clémençon et al. (2008) (page 846) that writes the bipartite ranking risk in its equivalent form

$$L(s) = \mathbb{P}\{(\zeta - \zeta')(s(\mathbf{Z}) - s(\mathbf{Z}')) < 0\} \,. \tag{2.2.14}$$

Let a bivariate ranking rule $r_s : \mathscr{X} \times \mathscr{X} \to \{-1, 1\}$ depending on the scoring function s defined by the equivalence

$$r_s(z, z') = 1 \quad \textit{iff} \quad s(z) \geq s(z') \,, \tag{2.2.15}$$

and supposing the absence of ties or else $r_s(z, z) = 0$. By considering the label $\tilde{\zeta} = (\zeta - \zeta')/2$ for a given pair $(\mathbf{Z}, \mathbf{Z}')$, the bipartite ranking risk can be formulated as a pairwise classification loss $L_r(s) = \mathbb{P}\{\tilde{\zeta} \times r_s(\mathbf{Z}, \mathbf{Z}') < 0\}$. The optimal ranking error is proved to be

$$L_r^* = \frac{1}{2p(1-p)} \mathbb{E}[\min(\eta(\mathbf{Z})(1 - \eta(\mathbf{Z}')), \eta(\mathbf{Z}')(1 - \eta(\mathbf{Z})))] \,, \tag{2.2.16}$$

while the excess of risk to

$$L_r(s) - L_r^* = \mathbb{E}[|\eta(\mathbf{Z}) - \eta(\mathbf{Z}')| \times \mathbb{I}\{r_s(\mathbf{Z}, \mathbf{Z}') \times (\eta(\mathbf{Z}) - \eta(\mathbf{Z}')) < 0\}] \,, \tag{2.2.17}$$

see Clémençon et al. (2008) for additional results on uniform error bounds and fast rates guarantees. This construction is at the heart of state-of-the-art algorithms related to learning-to-rank problems, as will be detailed in the sequel. The basic idea is to treat each of all possible pairs formed as a single variable. Nevertheless, this is insufficient to reduce the algorithmic complexity as it remains quadratic in the number of pairs. Algorithmic procedures were proposed to attempt a reduction in complexity. Ailon and Mohri (2008) proposed algorithmic tricks to reduce the complexity for pairwise classification loss, from $\mathscr{O}(N^2)$ to $\mathscr{O}(N \log(N))$ ($\mathscr{O}(k \log(N + k))$ if top-$k \ll N$ instances considered).

Another possibility, largely and implicitly followed in the literature, is using computational tricks or learning the discriminant function among simple classes, *e.g.*, linear forms. However, when looking at classic methods related to univariate losses, for instance in classification, sate-of-the-art algorithms imply convex surrogate loss functions (*e.g.* boosting, support vector machines). Apart from inheriting powerful results from the convex optimization theory, it occurs to be a very useful trick, circumventing the high complexity of simple problems. Indeed, as argued by Devroye et al. (1996) Section 4.6, the minimization related to the $0-1$ loss can quickly become NP hard. This typical

method aims to obtain an objective function admitting at least one minimum and intrinsically accelerating the order of convergence; fundamental examples are gathered in Table 2.1.

Loss	Formula
binary	$\ell(s,z',z) = \mathbb{I}\{z' > z\} + (1/2)\mathbb{I}\{z' = z\}$
squared	$\ell(s,z',z) = (1 - s(z') + s(z))^2$
logistic	$\ell(s,z',z) = \log(1 + e^{s(z')-s(z)})$
exponential	$\ell(s,z',z) = e^{s(z')-s(z)}$
hinge	$\ell(s,z',z) = \max(0, 1 - (s(z') - s(z)))$

Table 2.1. Bivariate margin losses associated to the bipartite ranking risk.

The classic algorithms applied to bipartite ranking and more broadly to learning-to-rank methods are based on this approach. For instance, `RankBoost` (Freund et al. (2003)) is an extension of the `AdaBoost` (Freund and Schapire (1997)) leading to a bivariate exponential loss, selecting the *weak* ranker corresponding to the largest decrease in the loss function. In Joachims (2006); Rakotomamonjy (2004), Support Vector Machines (SVM) are adapted to `RankSVM`, minimizing the bipartite risk with surrogate hinge loss. Also, `RankNet` and `LambdaRankNet`, introduced by Burges et al. (2005), are adaptation of Neural Nets (NN) by optimizing the binary cross entropy loss with a modification of the backpropagation step. See also Narasimhan and Agarwal (2017) for SVM-based algorithm optimizing the partial AUC (defined in Appendix section B). However, these algorithms are implicitly derived to rely on linear scoring functions, such that $r_s(z,z') = f_s(z-z')$, where f_s is a univariate loss function, thus drastically simplifying their implementation. For instance, the complexity of linear `RankSVM` with L_1-loss reduced the quadratic complexity to at least $\mathcal{O}(N\tilde{n} + N\log(N))$, with $\tilde{n}$ the average number of non-zero features per observation, plus a loglinear term depending on the optimization algorithm, see *e.g.* Joachims (2006). More generally, we refer to Menon and Williamson (2016) for a comprehensive review on the bipartite ranking problem and all its related formulations.

Optimizing in the ROC space. Another perspective that motivates this manuscript, is inherited by relation (2.2.11) where the order induced by a given scoring function is analyzed thanks to the ROC curve. This approach leads to more interpretable versions of the risk, as the AUC sums the number of pairs for which the positive instances are ranked higher than the negative ones. In this sense, many alternative loss functions exist, while the aforementioned algorithms aim to minimize a related version to the empirical AUC. First, a method extending tree-based methods was proposed by Clémençon and Vayatis (2009b); Clémençon et al. (2011), named `TreeRank`. Formulated to maximize the AUC at each step (greedy), it learns an optimal piecewise constant scoring function by approximating the ROC curve at each step using the 'temporary' optimal scoring function. In particular (applied) problems, misranking a 'positive' instance with a low score do not have the same interpretation/impact as a high score. In this sense, many works have introduced different cost-sensitive functions that emphasize the observations with higher ranks. In Järvelin and Kekäläinen (2000), authors introduced the discounted cumulative gain (DCG) factor, where the loss is weighted by the scores, whereas Boyd et al. (2012); Clémençon and Vayatis (2007) focused on only to the *best* instances, *i.e.*, those falling in the quantile of the whole mapped sample. Rudin (2006) and later Agarwal (2011) proposed a smooth loss function named p/infinite-norm push, with $p > 0$, that flattens the effect of the low ranks by 'pushing' the 'negative' instances far from the 'positive' ones. It was related to a generalization of `RankBoost` in those works. Lastly, cost-sensitive functions can recover the $0-1$ loss functions for particular choices of parameters (cost and threshold), as studied

in Krzanowski and Hand (2009). These alternatives will be detailed in Chapters 5 and 6.

Relating bipartite ranking to univariate binary models. However these two directions, the complexity remains high for pairwise algorithms (of order $\mathcal{O}(N^2)$) as the obtained statistics are non-*i.i.d.* sums of highly correlated terms. We discuss on a series of works bridging the gap to univariate formulations and, more precisely, by looking into the relation between optimal elements when writing the bipartite ranking loss as univariate functional. Promising empirical results using binary classifiers as ranking classifiers are obtained, when compared to bipartite ranking rules for state-of-the-art algorithms. For instance, Rudin and Schapire (2009) compared ranking performances of `RankBoost` and `AdaBoost`, both algorithms were shown to be equivalent, when one chooses the exponential loss, see Theorem 10. In particular, they proved that if the constant scoring function is included in the set of the weak learners of `AdaBoost`, then it converges precisely to the same AUC as `RankBoost`, under explicit necessary conditions. However, their relation is far from being straightforward. Simple examples can show that there is not necessarily equivalence between a *good* bipartite ranking and a *good* binary classification in the sense of the minimization of the risk.

These experiments motivated fundamental contributions and formalization relating bipartite ranking models to binary classification (BC) model, learning the binary label, and to class probability estimation (CPE) model, learning the probability of an observation to belong to a class. The purpose is to explicit the necessary conditions to bound the corresponding losses when transferring one model to another. In particular, the comparison of their excess of risk allows such transfers, which is particularly studied in Agarwal (2014); Clémençon and Robbiano (2011); Ertekin and Rudin (2011); Narasimhan and Agarwal (2013); W. Kotlowski (2011). Figure 2.2 summarizes these results by supplementing the work of Narasimhan and Agarwal (2013). The optimal element/solution of a model can be used as classifier/scorer for another. It largely democratizes bipartite ranking problems by drastically reducing their computational difficulty. We qualify a *weak* (W) model transfer if the mapping depends on the underlying probability distribution of the problem, thus implying a data-driven procedure in practice. On the contrary, *strong* (S) model transfer asserts that one only needs to threshold to 0.5. For both types/characterizations, if a statistically consistent estimator is obtained for one problem then it is for the other one.

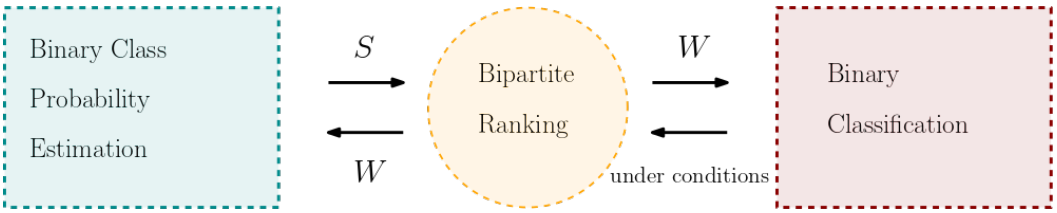


Figure 2.2. Equivalence relations for bipartite ranking to binary class probability estimation and to binary classification. *W* refers to weak model transfer, *S* refers to strong model transfer.

First, one obtains a consistent ranking rule with a consistent CPE estimator that approximates the prior probability $\eta(z)$, as it incidentally corresponds to a plug-in estimator. This result was formalized in Clémençon and Robbiano (2011), Proposition 4 therein. Then, transferring from BR to CPE model requires selecting a threshold minimizing the corresponding expected error (depending on the distribution) or its empirical counterpart (depending on the random sample). Hence, the reciprocal is relatively straightforward by simply transforming the rank to a probability *via* introducing a monotone map minimizing the binary CPE error on the sample.
Also, transferring the optimal scorer in the sense of BR for using it as a classifier for a cost-sensitive

$0-1$ loss is proved by Narasimhan and Agarwal (2013), Theorem 6 therein, for which a sketch of a possible two-stage transfer procedure is proposed, see Remark 7. First, the observations are ranked/ordered. Then, the threshold that determines if a positive or negative label is attributed to an instance is learned based on the sample.
Transferring BC to BR has a high impact on a large number of theoretical guarantees and algorithms that have been established in the literature. An optimal binary classifier *w.r.t.* the $0-1$ loss cannot achieve good ranking performance. However, for balanced cost-sensitive losses, its excess of risk upperbounds the one of the BR model, see W. Kotlowski (2011). Nevertheless, it requires a prior estimate of $\mathbb{P}\{\zeta\}$ *i.e.* of the distribution of the labels to derive the balanced empirical risk. Fortunately, they also obtained similar results for two classic margin surrogates, such as exponential and logistic ones. Later, Agarwal (2014) generalized it to *strongly proper* (composite) losses, see Definition 7 and characterizations in Section 4. Briefly, this class includes typical examples such as the ones gathered in Table 2.1 (except for the binary loss), but also to (canonical) spherical and squared losses.The main theorem of Balcan et al. (2007) proves that, given a binary classification loss, the obtained bipartite ranking loss ($1-$AUC) is at most multiplied by 2.

2.2.3 An almost equivalent two-sample formulation

As argued in Section 2.2.1, the problem is completely defined by the knowledge of $(F,\ \eta)$, where in particular, the sample $(\mathbf{Z},\ \zeta)$ is considered conditioned on the *r.v.* ζ for the bipartite risk error. In fact, a two-sample formulation can be equivalently defined *w.r.t.* the latter risk as follows. Let G (*resp.* H) the marginal distribution of $\mathbf{Z} \mid \{\zeta = 1\}$ (*resp.* $\mathbf{Z} \mid \{\zeta = -1\}$) and $p = \mathbb{P}\{\zeta = 1\}$. The likelihood ratio boils down to $\Psi(z) = dG/dH(z)$ and one can consider the triplet $(G,\ H,\ p)$ to completely define the probabilistic problem. This formulation corresponds to the case where the labels are deterministic. In this new setting, F is defined as the mixture *c.d.f.* of the two independent *r.v.* $\mathbf{X} \sim G$ and $\mathbf{Y} \sim H$, with p being the 'theoretical' proportion of the $\mathbf{X}$s among the pooled sample, yielding to $F = pG + (1-p)H$. In particular, for a given scoring function $s \in \mathscr{S}$, the *c.d.f.* of $s(\mathbf{X})$ (*resp.* $s(\mathbf{Y})$) is defined by G_s (*resp.* H_s), refer to Definition 9. The expected loss of Definition 7 simply is

$$L(s) = \mathbb{P}\{s(\mathbf{X}) > s(\mathbf{Y})\} + \frac{1}{2}\mathbb{P}\{s(\mathbf{Y}) = s(\mathbf{X})\} =: 1 - \mathrm{AUC}_{\mathrm{H_s,G_s}}(\mathrm{s})\ . \tag{2.2.18}$$

As the expected risk written in (2.2.1) is the integral of the risk *w.r.t.* the conditional law on the labels, it is equivalent to the two-sample formulation (2.2.18) above. However the equivalence in expectation, the statistical version needs a precise procedure as follows. Chung and Romano (2013) introduced a coupling procedure, relating both modelings, in the more general setting of k-samples permutation statistical tests. Intuitively, the two models are not exactly equivalent as the vector of instances, say $\{\mathbf{X}_1, \ldots, \mathbf{X}_n, \mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$, with $n, m \in \mathbb{N}^*$ is not invariant by permutation, whereas $\{(\mathbf{Z}_1, \zeta_1), \ldots, (\mathbf{Z}_{n+m}, \zeta_{n+m})\}$ is. Precisely, the reciprocal formulation relies in the introduction of an auxiliary independent Bernoulli *r.v.* of parameter p, that determines at random the labels to attribute. We adapt it to our framework in Algorithm 2 below.

The output sample of Algorithm 2 does not necessarily contain all the observations from the initial samples, and the number of 'new' observations is random. While being practical, theoretical guarantees are established when applied to permutation tests in Chung and Romano (2013), proving that the asymptotic distribution of the new sample is asymptotically equal to the pooled sample from the $\mathbf{X}_i$s, $\mathbf{Y}_j$s, if $p - m/N = \mathscr{O}(N^{-1/2})$.

Algorithm 2: Coupling Argument: From the two-sample to the binary framework

Data: Independent *i.i.d.* samples $\{\mathbf{X}_i\}_{i\leq n}$ drawn from G and $\{\mathbf{Y}_j\}_{j\leq m}$ drawn from H.
Input: Theoretical proportion p, Bernoulli *r.v.* ε of parameter p.
Result: A *i.i.d.* sample $\{\mathbf{Z}_1,\ldots,\mathbf{Z}_N\}$ of distribution $F = pG+(1-p)H$.

```
Set i, j = 0, 0;
Step 1.
while i ≤ n or j ≤ m do
    Draw ε ∼ B(p) ;
    if ε = 1 then
        Set Z_{i+j} = X_i and i += 1;
    else
        Set Z_{i+j} = Y_j and j += 1;
    end
end
Step 2.
if i = n then
    Do Step 1: if ε = 1 sample from G until i + j = N;
end
else if j = m then
    Do Step 1: if ε = 0 sample from H until i + j = N;
end
Reorder the constructed sample by merging X_1,...,X_i, Y_1,...,Y_j with the sample at Step 2.
```

A statistical approach to the two-sample loss function. Define $n = \lfloor pN \rfloor$ and $m = N - p$, *s.t.* $n/N \to p \in (0,1)$, when $N \to +\infty$. The statistical version of the AUC can be estimated on the independent *i.i.d.* two samples $\{\mathbf{X}_1,\ldots,\mathbf{X}_n\}$ and $\{\mathbf{Y}_1,\ldots,\mathbf{Y}_m\}$, *resp.* drawn from G and H. First, define the empirical *c.d.f.* by $\widehat{G}_{s,n}(t) = (1/n)\sum_{i=1}^{n}\mathbb{I}\{s(\mathbf{X}_i)\leq t\}$, $\widehat{H}_{s,m}(t) = (1/m)\sum_{j=1}^{m}\mathbb{I}\{s(\mathbf{Y}_j)\leq t\}$, the empirical mixture distribution is incidentally

$$\widehat{F}_{s,N}(t) = (n/N)\widehat{G}_{s,n}(t) + (m/N)\widehat{H}_{s,m}(t)\ . \tag{2.2.19}$$

Hence, the empirical AUC based on these samples yields

$$\mathrm{AUC}_{\widehat{\mathrm{H}}_\mathrm{m},\widehat{\mathrm{G}}_\mathrm{n}}(\mathrm{s}) = \frac{1}{\mathrm{nm}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\sum_{\mathrm{j}=1}^{\mathrm{m}}\left(\mathbb{I}\{\mathrm{s}(\mathbf{Y}_\mathrm{j}) < \mathrm{s}(\mathbf{X}_\mathrm{i})\} + \frac{1}{2}\mathbb{I}\{\mathrm{s}(\mathbf{Y}_\mathrm{j}) = \mathrm{s}(\mathbf{X}_\mathrm{i})\}\right)\ . \tag{2.2.20}$$

This two-sample formulation will be studied at length throughout the manuscript (from Chap. 5 until 9), and we will show how the proposed generalization of R-statistics are scalar summaries of the ROC curve, just like the AUC(s) (2.2.20) as a particular case.

2.3 Anomaly ranking

This section introduces methods addressing the problem of detecting and ranking anomalies, known as anomaly ranking. While detecting anomalies has longstanding literature, ranking them is still a question at its early stage, and in particular if it is by means of ranking-based methods. Briefly, anomaly detection aims to identify from a set of observations, the ones that are qualified as abnormal, outliers, novelties, *i.e.*, that often correspond to rare occurrences. It is of particular interest in

applied fields such as in fraud and fault detection, network intrusion, monitoring systems in various organizations (health care for instance), and more generally for data processing purposes. Therefore, there is a plethora of articles, reviews, characterizations on such techniques, that can depend on the field of application. The following paragraphs provide an insight into particular concepts and aim to motivate the use of ranking methods, as proposed in Chap. 8. However, we do not provide a comprehensive review on anomaly detection methods as it goes beyond the scope of the manuscript. Frameworks range from supervised to semi/un-supervised, depending on whether both 'normal' and 'abnormal' data can be labeled. By definition, the major difficulty of detecting anomalies is based on their intrinsic structure: they represent only a very small proportion of the overall dataset, and hence are very *sparse* in the ambient space. This leads to highly imbalanced samples, for which learning a detector can naturally imply statistical bias or overfitting, or both, regardless of the approach. From now on, we place ourselves in point outlier detection, built to detect instances that are 'abnormal' with respect to the rest of the dataset.

Anomaly detection methods. The first statistical models considered that the underlying probability law of the 'normal' sample was known, usually set to Gaussian or Uniform if parametric, and defined anomalous the instances taking values 'far' from the distribution, *i.e.* greater than a threshold or lying in a quantile of fixed order. In other words, it aimed to determine how well the suspicious observations fitted the distribution. While being very natural, a major drawback is that this definition remains quite restrictive when facing high-dimensional real samples having unknown structure, refer for instance to the classic book of Barnett and Lewis (1994). In the late 90s, a great deal of approaches were inherited by data mining motivations, yielding to 'model-free' procedures. In particular, these relied on clustering models, where the abnormal instances are defined as those falling outside from the data-driven clusters, see *e.g.* Agrawal et al. (1998); Ester et al. (1996); Hinneburg and Keim (1998); Ng and Han (1994). Unfortunately, these methods are dependent on the method/algorithm and on the parameters of the clusters. Also, they yield to a binary notion of abnormality. Notice also that these algorithms aim to optimize the clusters, while ignoring the possible anomalies, whereas here the objective is to optimize the detection of anomalies.

Later, density-based approaches and algorithms refined the notion of abnormality that lead to 'ordering' the anomalies through either (i) a local characterization, supposing the 'outlier' to be within the global range but far from its neighbors; or (ii) a global one, where the 'outlier' is 'far' from the overall range. For the former, we refer to the recent review of Alghushairy et al. (2021) gathering such local methods for anomaly detection. In particular, Breunig et al. (2000) introduced the *Local Outlier Factor* (LOF) as *degree* of abnormality for each instance. Also, Wang et al. (2018) extended this notion through the *Connectivity-based Outlier Factor* (COF). Lastly, interesting algorithms lead for instance to: (i) peeling methods such as the Isolation Forest Liu et al. (2008), (ii) classification methods, see Bergman and Hoshen (2020) for semi-supervised modeling, Schölkopf et al. (1999) for one-class SVM, Steinwart et al. (2005) for plug-in techniques to binary classification unsupervised learning based on the estimation of the density level sets and using SVM. Ultimately, popular methods either model the anomaly detection problem as a one-class learning, or *via* two (highly) imbalanced domains/classes.

Anomaly ranking methods. We review anomaly ranking models rely on ranking-based learning techniques. Usually, the proposed methods are twofold: first an algorithm is performed to predict a nonbinary label for the studied sample, then a ranking criterion is used to order them by degree of abnormality. This can provide a way of interpreting these anomalies depending on the field of application. For instance, Dang et al. (2013) introduced a technique based on local dimensionality reduction via spatial projections. Müller et al. (2012) proposed a subspace clustering method such

that the ranking of the anomalies are fed by the multiple views through the subspaces. Also Müller et al. (2013) adapted to graph-based structures where the method ranks the nodes by they degree of deviation.

It is only recently, with the enthusiasm regarding the results of learning-to-rank algorithms and the gain in popularity of the ROC curves/analysis, that ranking methods are applied to anomaly detection, incidentally yielding to an intrinsic ordering of anomalies. This approach finds interest especially in industrial applications, where the ability to rank operations by degree of abnormality and prioritizing them can be highly time-consuming. Their motivations are twofold as they allow for: (i) nonparametric methods with possible high-dimensional complex data structure, (ii) explicit *degree of abnormality* through the concept of ranks. Hence, this approach provides a nonbinary *inlier/outlier* global characterization of the observations. In this sense, this category of methods are of particular interest for semi/un-supervised problems.

First, the bipartite ranking approach provides a very interesting framework for detecting the abnormal instances. Indeed, one can formulate this as building a *scoring function* $s : \mathscr{Z} \to]-\infty, \infty]$, such that a *good/optimal* one induces an ordering outlying a *quality* on $\mathscr{Z}$: the smaller the score is, the more probable the observation is anomalous and the lower its *rank* is. We briefly describe articles proposing different objective functions.

A simple application of ranking algorithms as described in the previous section, showed how to empirically interpret learned scores with pairwise loss functions in the context of bipartite ranking, see Carvalho et al. (2008). Following the classic ranking measures, a series of articles adapted learning-to-rank algorithms to anomaly ranking purposes. For instance, Frery et al. (2017) formulates the supervised anomaly ranking problem as to maximize the True Positive Rate among the top ranked instances, also known as average precision rate. It is an alternative approach to that of differentiating between normal and abnormal instances, for which the learning algorithm is based on stochastic gradient boosting optimization. Extensions have been studied, such as Lamba and Akoglu (2019) for online anomaly ranking model.

A closely related notion is inherited by the works of Einmahl and Mason (1992); Polonik (1997) that developed the concept of *minimum volume set*, defining the threshold to split the spatial regions wherein a multivariate random variable $\mathbf{Z}$ of distribution F valued in the measurable space $\mathscr{Z} \subset \mathbb{R}^d$, with $d \geq 1$, takes values with low or high probability. The idea is to define the *quality* of a sample, through its sparsity (*e.g.* its spread) in the feature space. This probability level is fixed and defines the threshold for considering a region as abnormal. The aim is to find the optimal set $\Omega^*(\alpha)$ of mass at least $\alpha \in (0,1)$ such that

$$\min_{\Omega \subset \mathscr{Z}} \lambda(\Omega) \quad s.t. \quad \mathbb{P}\{\mathbf{Z} \in \Omega\} \geq \alpha \,, \tag{2.3.1}$$

where λ is the Lebesgue measure on $\mathscr{Z}$ and Ω measurable subset of $\mathscr{Z}$. In fact, *Mass Volume* (MV) curves extend this definition by plotting all the possible thresholds in $(0,1)$ as a Probability-Measure plot, when redefining the problem *w.r.t.* the class of scoring functions as follows.

Definition 14 (Definition 2, Clémençon and Jakubowicz (2013))**.** *Let a class of scoring functions* $\mathscr{S} = \{s : \mathscr{Z} \to \mathbb{R} \cup \{+\infty\},\ s \ \textit{measurable}\}$. *The Mass Volume (*MV*) curve of a scoring function s is the parametric curve:*

$$\mathrm{MV_s} : \mathrm{t} \in \mathbb{R} \mapsto (1 - \mathrm{F_s(t)}, 1 - \lambda_\mathrm{s}(\mathrm{t})) \,,$$

valued in $[0,1]^2$*, where for all* $t \in \mathbb{R}$*,*

$$\begin{aligned} F_s(t) &= \mathbb{P}\{s(\mathbf{Z}) \leq t\} \\ \lambda_s(t) &= \lambda(\{z \in \mathscr{Z},\ s(z) \leq t\}) \,. \end{aligned}$$

The problem is to find the optimal function s^* such that at fixed level α, $\Omega_s = \{z \in \mathscr{X},\ s(z) \geq t\}$ is solution of (2.3.1) over the class of scoring functions, for which the optimal elements are shown to satisfy the following result.

Proposition 15 (Proposition 3, Clémençon and Jakubowicz (2013))**.** *Let* $\mathbf{Z}$ *absolute continuous of bounded d.f.* $f(z)$*, such that* f *has no flat parts a.e. The class of optimal scoring functions is defined as the set:*

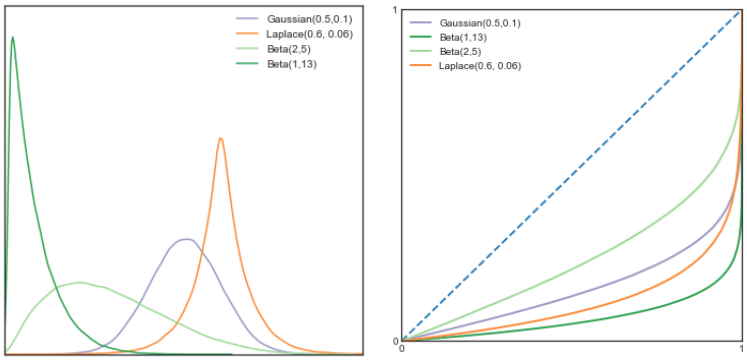
$$\mathscr{S}^* = \left\{ s \in \mathscr{S} \ \ s.t.\ for\ all\ z,\ z'\ in\ \mathscr{Z} :\ \ f(z) < f(z') \Rightarrow s^*(z) < s^*(z') \right\} . \tag{2.3.2}$$

And for all $s^* \in \mathscr{S}^*$*, the optimal* MV *curve satisfies:*

$$\mathrm{MV}^*(\alpha) \leq \mathrm{MV}_s(\alpha), \quad \forall \alpha \in (0,1) , \tag{2.3.3}$$

where in particular $\mathrm{MV}^* = \mathrm{MV}_f$.

In particular, by mimicking the density function f, an optimal scoring function thus induces an ordering for the observations depending on their *quality*: the smaller the score of an observation is, the more probable it is anomalous. Therefore the MV curve is a functional criterion for measuring the quality of a scoring function in the sense of the minimization of the mass volume set of (2.3.1), for all fixed levels α, as illustrated in Fig. 2.3.



a. Probability distributions b. Corresponding MV_f curves

Figure 2.3. Examples of distributions and their related MV curves.

This approach leads to a series of works, with in-depth theoretical analysis, formulated at first as unsupervised one-class framework (see Clémençon and Jakubowicz (2013)) and related to statistical analysis of extreme regions (see Goix et al. (2015); Thomas et al. (2017)), resulting to its generic formulation in Clémençon and Thomas (2018); Thomas et al. (2017). The generalization performance of the optimal elements are derived at length as well as the estimation of confidence region, accompanied with a smooth consistent bootstrap procedure. In addition, one can notice its intimate relation to ROC analysis, and in particular in supervised frameworks, see Clémençon and Robbiano (2014).

Lastly, neighbor ranking approaches are developed, aiming to estimate a local outlier rank-based criterion, that learns to rank the instances *w.r.t.* its proximity degree to its neighbors, see for instance Huang et al. (2013) for a Rank based Detecting Algorithm (RBDA), but also Bhattacharya et al. (2015); Huang et al. (2011); Qian et al. (2014). Finally, Perini et al. (2020) introduced a ranking method to measure the robustness of anomaly detection measures.

3 | Some Concentration Results

Abstract. This chapter outlines fundamental concepts at the crossroads of probability theory in Banach spaces and mathematical statistics, particularly for the (non)-asymptotic analysis of semi/non-parametric models. As foreshadowed in the previous chapter, while constructing a risk functional for a given model can be natural, choosing the *best* class of functions over which to optimize the loss, or at least the one having *good* properties, is of great difficulty. Indeed, in practice, the algorithm should output an empirical solution that converges to the oracle to minimize the risk among a class of test functions. This chapter provides probabilistic tools that allow for nonasymptotic concentration bounds, of mainly exponential form, to analyze the fluctuations of the empirical measure *w.r.t.* the true one. In particular, *good* properties of a possibly infinite class of functions will be detailed, and referred to as its *complexity*. More generally, if uniform deviation bounds are proved, it is possible to control the excess of risk of the chosen model, and generalization guarantees for the empirical solution are obtained. While motivating these results with the Empirical Risk Minimization (ERM) theory, such (uniform) inequalities will be recalled, especially when it is possible to quantify the *complexity* (*e.g.* entropy measures, combinatorial counting, *etc.*) of the class of functions on which it is indexed. In the last part, similar results for higher-order statistics, known as U-statistics, will be adapted to the current setting of the manuscript. Incidentally, fundamental decomposition techniques leveraging these statistics will be detailed.

Contents

3.1 Motivation: an introduction to Empirical Risk Minimization

Learning theory *"is posed as a problem of function estimation"* with either local or global solutions, as introduced by Vapnik (1992). This chapter details results on global solutions through the classic framework of Empirical Risk Minimization (ERM) as a particular formulation of M-estimation, which are key for understanding the analysis of infinite collections of R-statistics. When based on random samples of independent observations, we provide results for controlling the nonasymptotic efficiency of non/semi-parametric models, thanks to *concentration inequalities*. These establish bounds on the fluctuations of functionals of independent random variables around their expectation/mean. Also, they generalize confidence bounds and establish relations between the intrinsic parameters to the model and the size of the samples.

3.1.1 Problem formulation

Consider two input/output *r.v.* $(Z,\ \zeta)$, defined on the probability space $(\Omega,\ \mathscr{A},\ \mathbb{P})$ and taking their values in a measurable space $\mathscr{Z}\times\Xi$, where the probability distribution P is unknown. A learning algorithm aims to estimate the dependence relation between the input and output variables that is modeled by a function $f:\mathscr{Z}\to\Xi$ of a class $\mathscr{F}$. The *risk/error* committed is quantified by means of a *loss* function, say $\ell:\Xi\times\Xi\to\mathbb{R}$, averaged *w.r.t.* the underlying distribution defined as the functional

$$\mathscr{R}_\ell(f)=\mathbb{E}[\ell(f(Z),\ \zeta)]\,, \tag{3.1.1}$$

for all $f\in\mathscr{F}$. The *optimal* function f^* is defined as the minimizer of the risk such that $\mathscr{R}_\ell(f^*)=\inf_{f\in\mathscr{F}}\mathscr{R}_\ell(f)=:\mathscr{R}_\ell^*$.

Example 16. (BINARY CLASSIFICATION) *One of the most studied models is the binary classification, aiming to learn the classifier f that labels the observations by 0 or 1 (or equivalently by -1 or $+1$). For instance, when associated to the binary loss (or known as the $0-1$ loss), defined as $\ell_{01}:(x,y)\in\mathscr{Z}\times\{0,1\}=\mathbb{I}\{f(x)\neq y\}\in\{0,1\}$, the expected risk corresponds to the probability of mislabeling an observation given a function f. It equals to $\mathscr{R}_{01}(f)=\mathbb{P}\{f(Z)\neq\zeta\}$ and is minimized by the classifier $f^*(z)=\mathbb{I}\{\eta(z)>1/2\}$, based on the posterior probability $\eta(z)=\mathbb{P}\{\zeta=1\,|\,Z=z\}$.*

Solving this optimization problem requires the knowledge of P, leading to considering its empirical counterpart instead. Suppose a random sequence $(Z_1,\ \zeta_1),\dots,(Z_N,\ \zeta_N)$ is observed, of $N\in\mathbb{N}^*$ independent copies drawn from P. Therefore, the associated empirical minimizer of the expected risk is defined by

$$\hat{f}_N \in \underset{f\in\mathscr{F}}{\arg\min}\, \widehat{\mathscr{R}}^{\ell}_N(f) = \underset{f\in\mathscr{F}}{\arg\min}\, \frac{1}{N}\sum_{i=1}^{N} \ell(f(Z_i),\, \zeta_i)\ . \tag{3.1.2}$$

The definition of the expected risk and all the more of its empirical counterpart highlight the importance of the choice of $\mathscr{F}$ *w.r.t.* the law P. Indeed, the whole class $\mathscr{F}$, that includes the oracle function f^* if it exists, is unknown in practice and usually a subset in considered, say $\mathscr{F}_0 \subset \mathscr{F}$ (in the sense that $\mathscr{F}$ is the universe/perfect knowledge). One can only hope that f^* yields or can be approximated in the chosen subclass. Based on the statistical sample, the *quality* of a predictor f can be defined *via* its performance measure through the *excess of risk*

$$\mathscr{E}_\ell(f) = \mathscr{R}_\ell(f) - \mathscr{R}_\ell(f^*)\ . \tag{3.1.3}$$

However, in the most generic formulation, one can consider the minimizer $h^*_{\mathscr{F}_0}$ on the chosen subclass such that the decomposition holds

$$\mathscr{E}_\ell(\hat{f}_N) = \underbrace{\mathscr{R}_\ell(\hat{f}_N) - \mathscr{R}_\ell(h^*_{\mathscr{F}_0})}_{\textbf{estimation error}} + \underbrace{\mathscr{R}_\ell(h^*_{\mathscr{F}_0}) - \mathscr{R}_\ell(f^*)}_{\textbf{approximation error}}\ . \tag{3.1.4}$$

In fact, the estimation error can be upperbounded as follows, revealing the deviation of the empirical risk *w.r.t.* its mean that need to be controlled and analyzed.

Lemma 17. *Let $\mathscr{F}_0 \subset \mathscr{F}$. The estimation error related to the loss function ℓ and based on the random i.i.d. sample $(Z_1,\, \zeta_1),\ldots,(Z_N,\, \zeta_N)$ is bounded by:*

$$\mathscr{E}_\ell(\hat{f}_N) \leq 2\, \sup_{f\in\mathscr{F}_0} |\, \mathscr{R}_\ell(f) - \widehat{\mathscr{R}}^{\ell}_N(f)\, | + \mathscr{R}_\ell(h^*_{\mathscr{F}_0}) - \mathscr{R}_\ell(f^*)\ . \tag{3.1.5}$$

PROOF.

$$\begin{aligned}
\mathscr{R}_\ell(\hat{f}_N) - \mathscr{R}_\ell(f^*) &= \mathscr{R}_\ell(\hat{f}_N) - \widehat{\mathscr{R}}^{\ell}_N(\hat{f}_N) + \widehat{\mathscr{R}}^{\ell}_N(\hat{f}_N) - \widehat{\mathscr{R}}^{\ell}_N(f^*) + \widehat{\mathscr{R}}^{\ell}_N(f^*) - \mathscr{R}_\ell(f^*) \\
&\leq 2\, \sup_{f\in\mathscr{F}_0} |\, \mathscr{R}_\ell(f) - \widehat{\mathscr{R}}^{\ell}_N(f)\, |
\end{aligned}$$

□

Lemma 17 highlights that the suboptimality of $\hat{f}_N$, in the sense of the excess of risk, is controlled by the uniform fluctuations of $(\mathscr{R}_\ell - \widehat{\mathscr{R}}^{\ell}_N)$ over the possible infinite class $\mathscr{F}_0$. Secondly, it bounds the error when estimating the expected risk $\mathscr{R}_\ell(\hat{f}_N)$ by its statistical version $\widehat{\mathscr{R}}^{\ell}_N(\hat{f}_N)$. The following example illustrates a model where it is straightforward to bound the expected risk of the empirical estimator, under very restrictive assumptions.

Example 18. (BINARY CLASSIFICATION) *Considering Ex.* (16)*, if both distributions of the observations conditionally on the labels are separable for an element of $\mathscr{F}_0$ and $\mathscr{F}_0$ is supposed finite, then, one can show by the union bound that, for all $\varepsilon > 0$:*

$$\mathbb{P}\{\mathscr{R}_{01}(\hat{f}_N) > \varepsilon\} \leq |\mathscr{F}_0| e^{-N\varepsilon}\ ,$$

where $\hat{f}_N$ is the empirical minimizer of the empirical risk defined in Eq. (3.1.2)*.*

In the following, for a fixed loss function ℓ, the collection of estimators $(\mathscr{R}_\ell - \widehat{\mathscr{R}}^{\ell}_N)$ indexed by the class of functions $\mathscr{F}_0$, is defined as an *empirical process*. Precisely, when its uniform bound over $\mathscr{F}_0$ is measurable, we aim to control its fluctuations at a fixed level of probability, depending on the *complexity* of the class $\mathscr{F}_0$. In order to obtain these type of results, we will first walk trough concentration tail bounds for a fixed test function f, that are key to extend to the uniform norm over the whole class $\mathscr{F}_0$.

3.1.2 First concentration inequalities

In this section, mathematical tools are recalled as to bound in probability the fluctuation of $(\mathscr{R}_\ell - \widehat{\mathscr{R}}^\ell_N)$ for fixed loss ℓ and test f functions, depending on the sample size N. The first stated concentration inequality is from Hoeffding, considered the most elegant and simple way to control the sums of bounded random variables, with the only assumption of their mutual independence. Of course, refined inequalities can incorporate information on the moments of the *r.v.*, such as Bennett's (see Bennett (1962)) and Bernstein's (see Bernstein (1946)) inequalities.

Theorem 19 (Hoeffding tail inequality, Hoeffding (1963))**.** *Let $X_1, \ldots, X_N$, a sequence of $N \in \mathbb{N}^*$ independent r.v., s.t. $a_i \leq X_i \leq b_i$ a.s., with $(a_i, b_i) \in \mathbb{R}^2$, for all $i \leq N$. Then, for all $t > 0$,*

$$\mathbb{P}\left\{ \left| \sum_{i=1}^{N} (X_i - \mathbb{E}[X_i]) \right| \geq t \right\} \leq 2e^{-2t^2 / \Sigma_{i=1}^{N} (b_i - a_i)^2} . \tag{3.1.6}$$

To better understand what it encompasses, define $\delta = 2e^{-2t^2/\Sigma_{i=1}^{N}(b_i-a_i)^2}$. Then with probability at least $1-\delta$, it is possible to control almost surely the deviation of the sample mean *w.r.t.* its expectation by inverting Eq. (3.1.6) as follows

$$\frac{1}{N} \left| \sum_{i=1}^{N} (X_i - \mathbb{E}[X_i]) \right| \leq \sqrt{\sum_{i=1}^{N} (b_i - a_i)^2 \frac{\log(2/\delta)}{2N^2}} . \tag{3.1.7}$$

This bound expresses the importance of the spread effect for obtaining a *good* estimation of the expectation. It also provides an explicit bound, independent on the distribution of the sample, for which it is possible to exactly determine the sample size N required for the probabilistic control of the empirical bias.

Example 20. (BINARY CLASSIFICATION) *Considering the binary loss, and choosing $X_i = \mathbb{I}\{g(Z_i) \neq \zeta_i\}$, yields with probability at least $1-\delta$*

$$\left| \frac{1}{N} \sum_{i=1}^{N} (X_i - \mathbb{E}[X_i]) \right| \leq \sqrt{\frac{\log(2/\delta)}{2N}} . \tag{3.1.8}$$

The tail bound à la PAC is of order $\mathscr{O}_{\mathbb{P}}(N^{-1/2})$ that is classic for empirical estimators and processes, as will be shown is this manuscript.

In fact, Theorem 19 is a simple consequence of the following Hoeffding's inequality recalled below in Lemma 21, combined with Chernoff's bound illustrated in the subsequent proof, see Chernoff (1952). It is used in particular for the result Lemma 16 in Clémençon et al. (2021), proved in Chapter 4, Lemma 45 therein.

Lemma 21 (Hoeffding inequality, Hoeffding (1963))**.** *Let X a random variable s.t. $\mathbb{E}[X] = 0$ and $a \leq X \leq b$, $a, b \in \mathbb{R}$. Then, for all $s > 0$*

$$\mathbb{E}[e^{sX}] \leq e^{s^2(b-a)^2/8} . \tag{3.1.9}$$

PROOF. Let $t > 0$, $\lambda > 0$ and consider the centered *r.v.* $Z = \sum_{i=1}^{N} X_i - \sum_{i=1}^{N} \mathbb{E}[X_i]$, with $N \in \mathbb{N}^*$. Then, using sequentially Chernoff's bound, Hoeffding's Lemma and the independence of the X_is, one has

$$\begin{aligned}
\mathbb{P}\{Z \geq t\} &= \mathbb{P}\left\{e^{\lambda Z} \geq e^{\lambda t}\right\} \\
&\leq e^{-\lambda t}\mathbb{E}[e^{\lambda Z}] \\
&\leq e^{-\lambda t}\prod_{i\leq N}\mathbb{E}[e^{\lambda(X_i-\mathbb{E}[X_i])}] \\
&\leq \exp\left\{-\lambda t + \lambda^2 t^2 \sum_{i=1}^{N}(b_i-a_i)^2/8\right\} \\
&\leq \inf_{\lambda>0}\exp\left\{-\lambda t + \lambda^2 t^2 \sum_{i=1}^{N}(b_i-a_i)^2/8\right\} .
\end{aligned}$$

The bound is obtained with the optimal parameter $\lambda^* = 4t/\sum_{i=1}^{N}(b_i-a_i)^2$. □

Since these results, improvements for exponential bounds have been obtained for independent functions of random variables, first thanks to martingales methods (see McDiarmid (1989, 1998)), then to information-theoretic methods (see Ledoux (2001), Chapter 6) and also to induction methods (see Talagrand (1996a, 1995, 1996b)). We refer to Boucheron et al. (2005) for a thorough study of these methods and applications. Nevertheless, we present McDiarmid's bounded difference inequality for its great adaptability to statistics and processes that will be used in the proofs of Chapter 5.

Definition 22. *Let $\mathcal{X}$ a set and $f : \mathcal{X}^N \to \mathbb{R}$ a measurable function of N variables. The function f satisfies the bounded difference inequality, if for the real constants $c_1,\ldots,c_N$ and for all $i \leq N$, we have*

$$\sup_{x_1,\ldots,x_N,x_i'\in\mathcal{X}} |f(x_1,\ldots,x_N) - f(x_1,\ldots,x_i',\ldots,x_N)| \leq c_i . \tag{3.1.10}$$

Given such a function f, McDiarmid proved the following exponential tail bound, see McDiarmid (1989, 1998).

Lemma 23 (McDiarmid bounded difference inequality, McDiarmid (1989, 1998))**.** *Let $X_1,\ldots,X_N$ a sequence of independent r.v. valued in $\mathcal{X}$. Consider a function f satisfying the bounded difference inequality, with constants $c_1,\ldots,c_N \in \mathbb{R}$. Define $Z = f(X_1,\ldots,X_N)$, then for all $t > 0$*

$$\mathbb{P}\{|Z-\mathbb{E}[Z]| \geq t\} \leq 2e^{-2t^2/C} , \tag{3.1.11}$$

where $C = \sum_{i=1}^{N} c_i^2$.

These fundamental concentration tools are essential to understanding the quantification of the *quality* of an estimator when analyzing its fluctuations *w.r.t.* its mean. But also, it is necessary as to analyze its uniform fluctuations over possibly infinite classes of functions. The following section develops techniques for this last point.

3.2 Empirical processes

This section focuses on constructing concentration inequalities for uniform bounds of empirical processes, *i.e.* , for collections of statistics indexed by *infinite* classes of functions. Typically, it is of particular interest in ERM as it allows for bounding in probability the fluctuations of

$$\sup_{f\in\mathcal{F}} |\, \mathscr{R}_\ell(f) - \widehat{\mathscr{R}}_N^\ell(f) \,| \ .$$

Two key concepts for measuring the complexity of functional classes are presented: the uniform covering numbers and the *Vapnik-Chervonenkis* (VC)-dimension. Historically, the former was introduced by A. Kolmogorov in the 1950s, in the context of new metric approaches for mathematical analysis with the development of covering numbers, entropy metrics, *etc.* Whereas the latter results from combinatorial theory, proposed by Vapnik and Chervonenkis (2015) (in fact in 1968). This new method defined the complexity of collections of subsets. It led to a characterization of the complexity of an infinite class of functions when associated with the Empirical Process theory. In fact, both characterizations are intimately related. In particular, these notions are essential to the control of uniform deviations of such processes called *uniform generalization bounds*. Notice that recently, the works of Bartlett and Mendelson (2003); Koltchinskii and Panchenko (2000) provided fundamental advances, which introduced a Gaussian complexity to the continuous models named Rademacher complexity. It allows for an alternative concept for measuring the risk associated with infinite classes of functions.However, this manuscript only considers combinatorial and entropy approaches.

Notation. Throughout this section, $\mathscr{F}$ is a class of measurable functions $f : \mathscr{X} \to \mathbb{R}$, $\|\mu\|_{\mathscr{F}} = \sup_{f\in\mathscr{F}} |\mu(f)|$ the sup norm of $\mathscr{F}$ *w.r.t.* a measure μ of $(\Omega, \mathscr{A}, \mathbb{P})$.

3.2.1 Measuring the complexity of classes of functions

Given a sequence of *i.i.d. r.v.* $X_1, \ldots, X_N$, with $N \in \mathbb{N}^*$, defined on $(\Omega, \mathscr{A}, \mathbb{P})$ and valued in $\mathscr{X}$, the empirical measure μ_N based on the sample is classically defined by

$$\mu_N(A) = \frac{1}{N}\sum_{i=1}^{N} \delta_i(A) \;, \tag{3.2.1}$$

where $\delta_i(A) = \mathbb{I}\{X_i \in A\}$, for all $A \in \mathscr{A}$. From a statistical point of view, the empirical measure is the average of observations from a fixed sample that are valued in A. We first consider the uniform deviations of the empirical measure *w.r.t.* its expectation $\mu(A) = \mathbb{P}\{X \in A\}$, by

$$Z_N = \sup_{A\in\mathscr{A}} |\mu_N(A) - \mu(A)| \;. \tag{3.2.2}$$

We say that $\mathscr{A}$ is a *uniform Glivenko-Cantelli class*, if $\lim_{N\to\infty} \mathbb{E}[Z_N] = 0$. It follows that the map induced by μ_N on $\mathscr{F}$ is simply

$$f \mapsto \mu_N f = \int_{\mathscr{X}} f(x)\mu_N(dx) \;. \tag{3.2.3}$$

Under appropriate assumptions and for a fixed f, fundamental asymptotic/convergence properties can be applied to empirical processes, for instance, thanks to the law of large numbers and the central limit theorem, see the works Donsker (1952); Dudley (1999) and for a book reference Shorack and Wellner (2009) (initially published in 1986). We define below the centered and normalized empirical process related to the map (3.2.3).

Definition 24. *The centered and normalized empirical process indexed by $\mathscr{F}$, based on the i.i.d. sample $\{X_1, \ldots, X_N\}$, with $N \in \mathbb{N}^*$, is defined by the mapping:*

$$f \in \mathscr{F} \mapsto \frac{1}{\sqrt{N}}\sum_{i=1}^{N} (f(X_i) - \mathbb{E}[f(X)]) \;. \tag{3.2.4}$$

Example 25. (KOLMOGOROV-SMIRNOV TESTS.) *The statistics used for the Kolmogorov-Smirnov tests: either $(1/\sqrt{n})(\mu_N - \mu)$ or $\sqrt{nm}/\sqrt{n+m}(\mu_n - \nu_m)$, $N = n+m$, are empirical processes defined on the class of sets of the intervals of $\mathbb{R}$, i.e. $]-\infty, a]$, $a < \infty$. They are related resp. to goodness-of-fit and to two-sample univariate testing.*

Complexity of classes. In this manuscript and more generally in empirical process theory, we are interested in finding/defining a *good* class $\mathscr{F}$, so that such properties of uniform convergence over the whole class still hold. In particular, being able to characterize the complexity of a measurable function class *w.r.t.* empirical measure μ_N is a central point of this theory. The seminal example of classes composed of half-spaces of $\mathbb{R}$ (see Ex. 25) yields the empirical process related to the empirical distribution of X. Therefore, to extend these convergence properties over the whole class $\mathscr{F}$, we first present a natural way of measuring its complexity using the *entropy numbers*.

Definition 26. *Let $\varepsilon > 0$. The ε-entropy of a subset of normed space $(\mathscr{F}, \|\cdot\|)$, is the logarithm of the minimal number of balls of size ε i.e. $\{g,\ \|g-f\| \leq \varepsilon\}$, depending on the norm, to cover the whole class $\mathscr{F}$. The ε-covering number is defined by $N(\mathscr{F}, \|\cdot\|, \varepsilon)$ and the uniform entropy numbers as:*

$$\sup_{f \in \mathscr{F}} \log N(\mathscr{F}, \|\cdot\|, \varepsilon L_{\mathscr{F}}) \ ,$$

with $L_{\mathscr{F}} = \|F\|$ where F is the envelope function satisfying: $\forall x \in \mathscr{X}$, $\forall f \in \mathscr{F}$, $|f(x)| \leq F(x)$.

Notice, that usually $\mathscr{F}$ is a subset of L_p, with $p \geq 1$, *w.r.t.* a measure Q. Also, although the sequence of center functions need not belong to the class, they necessarily have a finite norm. Alternatively, the theory developed in Vapnik and Chervonenkis (2015) proposes a characterization of these types of classes and, in particular, of their *complexity* through a notion of combinatorial size/dimension.

Definition 27. *Let a collection $\mathscr{B}$ of subsets of a set Ω. A set $A \in \Omega$ is said to be shattered by $\mathscr{B}$, if the cardinality of $A \cap B = \{B \in \mathscr{B}, A \cap B\}$ equals to $2^{card(A)}$ i.e. if it picks out all the elements of A. The* VC*-index of $\mathscr{B}$ is the smallest $n \in \mathbb{N}$ for which no set of size/cardinality n is shattered by $\mathscr{B}$.*

For instance, by considering a n-sample $\{x_1, \ldots, x_n\}$, $\mathscr{B}$ shatters the n-tuple if each of the 2^n subsets *i.e.* expressed as $B \cap \{x_1, \ldots, x_n\}$, with $B \in \mathscr{B}$, can be picked. In particular, if the VC-index, $V > 0$, is finite then $\mathscr{B}$ is a VC-class of sets and Sauer's Lemma (see van der Vaart and Wellner (1996), Chapter 2.6) stated that the maximal number of subsets of a n-tuple picked out by a VC-class $\mathscr{B}$, is upperbounded by $\sum_{i \leq V} \binom{n}{i}$. Hence, a class of sets picks out no more than 2^n, for $V \leq n$. The following definition extends this combinatorial approach to classes of measurable functions by considering their subgraphs.

Definition 28. *Let $\mathscr{F}$ a class of measurable functions $f : \mathscr{X} \to \mathbb{R}$, the subgraph of f is the subset of $\mathscr{X} \times \mathbb{R}$ defined by $\{(x,t) \in \mathscr{X} \times \mathbb{R},\ f(x) > t\}$. If the class $\mathscr{F}$ is characterized by one of the two definitions below, it has finite* VC*-dimension V.*

1. *$\mathscr{F}$ is a* VC*-subgraph class (or* VC*-class) if the collection of all subgraphs of the functions of $\mathscr{F}$ forms a* VC*-class of sets (in $\mathscr{X} \times \mathbb{R}$).*
2. *$\mathscr{F}$ is a* VC*-major class if the sets $\{x :\ f(x) > t\}$, with f, t ranging in $\mathscr{F} \times \mathbb{R}$.*

Example 29. (BINARY CLASSIFICATION) *If the collection of classifiers $\mathscr{F}$ are binary i.e. valued in $\{0,1\}$ (see Ex. 16), then the class of subsets of a n-sample of observations picked out by $\mathscr{F}$, is included in $\{0,1\}^n$. Hence, if the* VC*-dimension V of $\mathscr{F}$ is finite, it is possible to shatter at most a sample of cardinality 2^V.*

The two concepts defined in 26 and 28 are to some extent intrinsically related, where for a given class, the uniform bound overall measures of its covering number can be upperbounded by a polynomial of the radius decreasing with V. In this line, the following characterization is considered for a general definition of bounded VC-type classes.

Definition 30. *A class $\mathscr{F}$ of real-valued functions defined on a measurable space $\mathscr{X}$ is a bounded* VC*-type class with parameters $(A, \mathscr{V}) \in (0, +\infty)^2$ and constant envelope $L_{\mathscr{F}} > 0$ if for all $\varepsilon \in (0,1)$:*

$$\sup_{Q} N(\mathscr{F}, L_2(Q), \varepsilon L_{\mathscr{F}}) \leq \left(\frac{A}{\varepsilon}\right)^{\mathscr{V}}, \tag{3.2.5}$$

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

Lastly, if $\mathscr{F}$ is a bounded VC-class with VC-dimension $V < +\infty$, the Eq. (3.2.5) is fulfilled with $\mathscr{V} = 2(V-1)$ and $A = (cV(16e)^V)^{1/(2(V-1))}$, where c is a universal constant, see *e.g.* Theorem 2.6.7 in van der Vaart and Wellner (1996).

Permanence properties. In van der Vaart and Wellner (1996), Chapter 2.6 gathers many typical examples of such classes. However, simple operations can extend those characterizations, guaranteed by permanence properties and gathered in Lemma 31 below.

Lemma 31 (Permanence properties, Lemmas 2.6.17-18, van der Vaart and Wellner (1996))**.** *Let $\mathscr{F}$ and $\mathscr{G}$ two* VC*-classes of functions on a set $\mathscr{X}$, $g : \mathscr{X} \to \mathbb{R}$, $\phi : \mathbb{R} \to \mathbb{R}$ monotone, then the following classes are* VC*-subgraphs:*

(i) $\mathscr{F} + g = \{f + g, \forall f \in \mathscr{F}\}$

(ii) $\mathscr{F} \cdot g = \{fg, \forall f \in \mathscr{F}\}$

(iii) $\phi \circ \mathscr{F}$

Also, if a collection of sets $\mathscr{B}$ is VC*-class, then the collection of indicators of sets in $\mathscr{B}$ is a* VC*-class of same index.*

Example 32. *By considering $\mathscr{G}$ a class of functions $g : \mathscr{Z} \subset \mathbb{R}^d \to \mathscr{X} \subset \mathbb{R}$, with $d \geq 1$, the class of indicator functions $\mathscr{F} = \{x \mapsto \mathbb{I}\{g(x) \geq 0\},\ g \in \mathscr{G}\}$ has a* VC*-dimension $V \leq d$.*

3.2.2 Uniform generalization bounds

This section briefly recalls classical contributions on exponential tail bounds for the uniform deviation between the empirical distribution and the true distribution function, starting with Dvoretzky-Kiefer-Wolfowitz inequality, see Dvoretzky et al. (1956).

Theorem 33 (Dvoretzky-Kiefer-Wolfowitz inequality, Dvoretzky et al. (1956))**.** *Consider $\mathscr{X} \subset \mathbb{R}$. Let $\mathscr{F}$ be the class indicator functions $\mathscr{F} = \{f : x \in \mathscr{X} \mapsto \mathbb{I}\{x \leq t\},\ t \in \mathbb{R}\}$, let $\{X_1, \ldots, X_N\}$ a n-sample drawn from F and of empirical distribution F_N valued in $\mathscr{X}$. Then, there exists a universal constant C, such that:*

$$\mathbb{P}\left\{\sqrt{N}\|\mu_N - \mu\|_\infty \geq t\right\} \leq Ce^{-2t^2}.$$

Later, Massart (1990) proved that the universal constant could be chosen equal to $C = 2$. Also, when considering broader classes of functions, Talagrand (1994, 1996b) presented significant developments of exponential tail bounds for the empirical measure, where $\mathscr{F}$ is of VC-type.

Theorem 34 (Talagrand inequality, Talagrand (1994)). *Let $\mathscr{F}$ a* VC*-type class of measurable functions defined on $\mathscr{X}$ and of constants $(A,\mathscr{V})$. Suppose $\mathbb{E}[\sup_{f\in\mathscr{F}}(1/N)\sum_{i\leq N}f^2(X_i)]\leq v$ and $\sup_{f\in\mathscr{F}}\|f\|_\infty\leq U$, with v, $U<\infty$. Then there exists a universal constant $K>0$, s.t. for all $t>0$:*

$$\mathbb{P}\{|\|\mu_N\|_{\mathscr{F}}-\mathbb{E}[\|\mu_N\|_{\mathscr{F}}]|\geq t\}\leq K\exp\left\{-\frac{Nt}{KU}\log\left(1+\frac{tU}{v}\right)\right\}.$$

Later, Giné and Guillou (2002) proposed a bound for the expectation of the uniform empirical measure, *s.t.* if combined with Talagrand's, a sharp exponential bound is obtained that will be used in some proofs of Chapter 5.

Theorem 35 (Giné and Guillou (2002)). *Let $\mathscr{F}$ a* VC*-type class of measurable functions defined on $\mathscr{X}$, and of constants $(A,\mathscr{V})$. Suppose $\mathbb{E}[\sup_{f\in\mathscr{F}}(1/N)\sum_{i\leq N}f^2(X_i)]\leq v$ and $\sup_{f\in\mathscr{F}}\|f\|_\infty\leq U$, with v, $U<\infty$. Then:*

$$\mathbb{P}\{\|\mu_N-\mu\|_{\mathscr{F}}\geq t\}\leq K\exp\left\{-\frac{Nt}{KU}\log\left(1+\frac{tU}{K(\sqrt{N}\sigma+U\sqrt{\log(AU/\sigma)})^2}\right)\right\},$$

as soon as

$$t\geq\frac{C}{N}\left(U\log(\frac{AU}{\sigma})+\sqrt{N}\sigma\sqrt{\log(\frac{AU}{\sigma})}\right).$$

3.3 U-processes

In the previous sections, the studied collections of statistics had the form of *i.i.d.*-sums of *r.v.*. This section extends those results to estimators taking the form of U-statistics, for which the theory goes back to the fundamental works of Halmos (1946) and Hoeffding (1948). As classic examples of high order statistics, the U-statistics are of particular interest as they encompass many nonparametric statistics. After introducing two forms of U-statistics, , we recall concentration results on their collections when indexed by classes of functions. While there is a rich literature on general forms of one-sample U-processes, in-depth results for multiple samples still lack.

3.3.1 U-statistics

Introduced by Halmos (1946) as the unique solution among the unbiased estimators of minimal variance, and by Hoeffding (1948) as tools for fundamental asymptotic theorems, the U-statistics are essential to the analysis and decomposition of R-statistics. To understand how they are built, P. R. Halmos proved that the U-statistics are solutions to the following generic problem, in the sense of the *Minimum-Variance Unbiased Estimator* (MVUE) class rendered, for instance, by the Theorem of Lehmann-Scheffé. Let θ be a functional defined on a set $\mathscr{F}$ of distribution functions of $\mathbb{R}$, such that

$$\theta:F\in\mathscr{F}\mapsto\theta(F).$$

Based on a *i.i.d.* sample $X_1,\ldots,X_N$ drawn from an unknown distribution F, the goal is to estimate the function $\theta(F)$. In its most generic formulation, the symmetric U-statistic solution of P.R. Halmos' problem is given by

$$U_n(\psi)=\binom{n}{q}^{-1}\sum_{\sigma\in\mathfrak{S}_n}\psi(X_{\sigma(1)},\ldots,X_{\sigma(q)}), \tag{3.3.1}$$

where $\mathcal{S}_n$ is the set of all permutations of $\{1,\ldots,n\}$, the function ψ is the kernel of the statistic and q its degree. Of course, symmetrical kernels are defined based on a measurable kernel $\tilde{\psi}$ by: $\psi(x_1,\ldots,x_q) = (q!)^{-1}\sum_{\sigma\in\mathcal{S}_q}\tilde{\psi}(x_{\sigma(1)},\ldots,x_{\sigma(q)})$. Two fundamental references gather results on U-statistics. Lee (1990) is the first comprehensive monograph on classic probability asymptotic theory and applications to statistical models. Korolyuk and Borovskich (1994), generalizes the results by relating/decomposing the U-statistics to/in reverse martingales valued in different types of spaces (Banach and Hilbert spaces). Additional references on U-statistics and extensions are *e.g.* Serfling (1980); Stute (1991).

Example 36. *Basic examples for estimating the parameters of a i.i.d. random sample $X_1,\ldots,X_n$, under some basic moment-based assumptions, and with X, $X' \overset{i.i.d.}{\sim} F$, are as follows:*

1. *mean: $\theta(F) = \mathbb{E}_{X\sim F}[X]$, then $U_n = (1/n)\sum_{i\leq n}X_i$*
2. *variance: $\theta(F) = \mathrm{Var}[X]$, then $U_n = \binom{n}{2}^{-1}\sum_{1\leq i<j\leq n}(X_i - X_j)^2$*
3. *covariance: $\theta(F) = \mathrm{Cov}[X,X']$, then $U_n = \binom{n}{2}^{-1}(1/2)\sum_{1\leq i<j\leq n}(X_i - X_j)(X'_i - X'_j)$*

Definition (3.3.1) highlights the knowledge that lies in this class of statistics thanks to the sum of correlated terms. Hence, the analysis developed for empirical processes is not applicable as such. A major contribution of Hoeffding is the decomposing of (3.3.1) in uncorrelated terms of variance with decreasing order, also known as the H-decomposition, see Hoeffding (1961). If the variables are square-integrable, the geometric interpretation generalizes the Hajek's projection, see Appendix section B.1 Lemma 92 therein. In particular, for $q = 2$, the first term of the decomposition equals to the projection statistic of Hájek's method.

Theorem 37 (Hoeffding decomposition, Theorem 1.6.1, Chapter 1, Lee (1990))**.** *Consider the U-statistic as defined in Eq. (3.3.1). For $j \in \{2,\ldots,q\}$, let $\psi_j(x_1,\ldots,x_j) = \mathbb{E}[\psi(x_1,\ldots,x_q) \mid X_1 = x_1,\ldots,X_j = x_j]$, such that the kernels are recursively defined:*

$$\begin{aligned} h^{(1)}(x_1) &= \psi_1(x_1) - \theta\ , \\ h^{(j)}(x_1,\ldots,x_j) &= \psi_j(x_1,\ldots,x_j) - \sum_{c=1}^{j-1}\sum_{\sigma\in\mathcal{S}_c} h^{(c)}(x_{\sigma(1)},\ldots,x_{\sigma(c)}) - \theta\ , \quad \forall j \in \{2,\ldots,q\}\ . \end{aligned}$$

Then the U-statistic of degree q is decomposed as

$$U_n(\psi) = \theta + \sum_{j=1}^{q}\binom{q}{j}U_n^{(j)}(h^{(j)}), \tag{3.3.2}$$

where the $U_n^{(j)}$ are the U-statistics based on kernel $h^{(j)}$ of degree j.

An important characterization of the kernels is their order of *degeneracy* (also related to the *rank* of the statistic). Indeed, a U-statistic is said to be *degenerate* or order c *w.r.t.* a probability measure, if the first c terms of the decomposition equal to zero *a.s.*. Its variance is of order $n^{-(c+1)}$. In particular, the order of degeneracy controls the limit distribution of the statistic, see *e.g.* Serfling (1980). The sequence $U_n^{(j)}$, for $j \leq q$ is of respective *rank* j such that the Theorem below holds true.

Theorem 38 (Theorem 1.6.2, Chapter 1, Lee (1990))**.** *Consider the result of Theorem 37. Then, for all $j \leq q$ and $c \leq j-1$,*

$$\mathbb{E}[h^{(j)}(X_1,\ldots,X_j) \mid X_1,\ldots,X_c] = 0 \tag{3.3.3}$$

and the kernels $\mathbb{E}[h^{(j)}(X_1,\ldots,X_j)] = 0$.

For clarity purposes the generic version of U-statistic was based on a unique *i.i.d.* sample. It is possible to define on r-samples, $r \geq 2$, of respective degree $(q_1, \ldots, q_r)$ with similar decomposition and known as *generalized* statistics, see Lee (1990) Chapter 2 therein. Nevertheless, in the context of two-sample R-statistics, solely nonsymmetric one-sample U-statistics of degree (2) and two-sample ones of degree $(1,1)$ are considered. We define these in the sequel with the same notations that will be used in Chapters 4, 5 and 6.

Definition 39 (One-sample U-statistics of degree 2)**.** *Let $n \geq 2$. Consider a i.i.d. sequence $X_1, \ldots, X_n$ drawn from a probability distribution μ on a measurable space $\mathscr{X}$ and $k : \mathscr{X}^2 \to \mathbb{R}$ a square integrable function w.r.t. $\mu \otimes \mu$. The one-sample U-statistic of degree 2 and kernel function k based on the X_i's is defined as:*

$$U_n(k) = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} k(X_i, X_j) \; . \tag{3.3.4}$$

As detailed, the statistic $U_n(k)$ is the MVUE of the parameter defined by $\theta(\mu) = \int k(x_1, x_2)\mu(dx_1)\mu(dx_2)$. In order exhibit the linear part of the statistic, the first term of its Hoeffding decomposition of $U_n(k)$ yields

$$\widehat{U}_n(k) = \frac{1}{n} \sum_{i=1}^{n} k_1(X_i) \; , \tag{3.3.5}$$

with $k_1 = k_{1,1} + k_{1,2}$ and for all $x \in \mathscr{X}$

$$\left\{ \begin{array}{ll} k_{1,1}(x) & = \mathbb{E}[k(X_1, x)] - \theta(\mu) \\ k_{1,2}(x) & = \mathbb{E}[k(x, X_2)] - \theta(\mu) \; , \end{array} \right.$$

while the degenerate part trivially equals to $U_n^{(2)} = U_n(k) - \theta(\mu) - \widehat{U}_n(k)$ and is of order $O_{\mathbb{P}}(1/n)$. Notice that $\widehat{U}_n(k)$ corresponds to the Hájek projection of $U_n(k) - \theta(\mu)$ onto the space of all random variables $\sum_{i=1}^n g_i(X_i)$ with $\int g_i^2(x)\mu(dx) < +\infty$, with the sequence g of measurable functions, as stated in Lemma 92, Appendix section B.1. In a same manner, we define below a two-sample U-statistics of degree $(1,1)$.

Definition 40 (Two-sample U-statistics of degree $(1,1)$)**.** *Let n, m in $\mathbb{N}^*$. Consider two independent i.i.d. sequences $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ respectively drawn from the probability distributions μ and ν on the measurable spaces $\mathscr{X}$ and $\mathscr{Y}$. Let $\ell : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$ be a square integrable function w.r.t. $\mu \otimes \nu$. The two-sample U-statistic of degree $(1,1)$, with kernel function ℓ and based on the X_i's and the Y_j's is defined as:*

$$U_{n,m}(\ell) = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \ell(X_i, Y_j) \; . \tag{3.3.6}$$

Similarly, $U_{n,m}(\ell)$ is the MVUE of $\theta(\mu,\nu) = \int \ell(x, y)\mu(dx)\mu(dy)$ and we exhibit the linear part of its Hoeffding decomposition, as follows

$$\widehat{U}_{n,m}(\ell) = \frac{1}{n} \sum_{i=1}^{n} \ell_{1,1}(X_i) + \frac{1}{m} \sum_{j=1}^{m} \ell_{1,2}(Y_j) \; ,$$

where for all $(x,y) \in \mathscr{X} \times \mathscr{Y}$, one has

$$\left\{ \begin{array}{ll} \ell_{1,1}(x) & = \mathbb{E}[\ell(x, Y_1)] - \theta(\mu,\nu) \\ \ell_{1,2}(y) & = \mathbb{E}[\ell(X_1, y)] - \theta(\mu,\nu) \; , \end{array} \right.$$

while the degenerate part equals to $U_{n,m}^{(2)}(\ell) = U_{n,m}(\ell) - \theta(\mu,\nu) - \widehat{U}_{n,m}(\ell)$ and is of order $O_{\mathbb{P}}(1/n) + O_{\mathbb{P}}(1/m)$. The Hájek projection of (3.3.6) is obtained by computing the orthogonal projection

of the recentered *r.v.* $U_{n,m}(\ell) - \theta(\mu,\nu)$ onto the subpace of L_2 composed of all random variables $\sum_{i=1}^n g_i(X_i) + \sum_{j=1}^m f_j(Y_j)$ with $\int g_i^2(x)\mu(dx) < +\infty$ and $\int f_j^2(y)\nu(dy) < +\infty$.

Example 41. *A classic example of two-sample U-statistic of degree $(1,1)$ is the Mann-Whitney statistic, of kernel $\ell(x,y) = \mathbb{I}\{y < x\} + (1/2)\mathbb{I}\{y = x\}$ on $\mathbb{R}^2$. It is a natural (unbiased) estimator of the* AUC*: when computed from univariate samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ with distributions G and H on $\mathbb{R}$, it is equal to $\mathrm{AUC}_{\widehat{H}_m,\widehat{G}_n}$ with the notations of Subsection B.3 and can be thus viewed as an affine transform of the rank-sum Wilcoxon statistic (2). The Hoeffding decomposition of the empirical* AUC *yields*

$$\mathrm{AUC}_{\widehat{H}_m,\widehat{G}_n} = \mathbb{P}\{X \geq Y\} + \frac{1}{n}\sum_{i\leq n}(\widehat{H}_m(X_i) - \mathbb{E}[H(X)]) - \frac{1}{m}\sum_{j\leq m}(\widehat{G}_n(Y_j) - \mathbb{E}[G(Y)]) + o_{\mathbb{P}_{G,H}}(1)\ . \tag{3.3.7}$$

The Hoeffding decomposition is the key to extend (limit) results known for *i.i.d.* averages (*e.g.* SLLN, CLT, LIL) to statistics of the type (3.3.6). In the technical analysis that are presented in Chapters 4 and 5, nonasymptotic uniform results are required for U-processes, namely collections of U-statistics indexed by classes of kernels. By means of the Hoeffding decomposition, concentration bounds for U-processes can be obtained by combining classic concentration bounds for empirical processes and concentration bounds for degenerate U-processes, such as the following section.

3.3.2 Concentration inequalities for degenerate one-sample U-processes

Similar to concentration bounds for empirical processes, this section encompasses concentration bounds for U-processes defined as collections of U-statistics indexed by classes of kernels. As formerly developed, the Hoeffding decomposition is key towards establishing various types of bounds. The study of a general U-statistic boils down to analyzing each uncorrelated term recursively on their degree, and particularly leads to degenerate statistics. Hence, a rich literature studies nonasymptotic guarantees of degenerate one-sample U-statistics. Remarkable results are, for instance, obtained by Arcones and Giné (1994) for exponential uniform bounds and asymptotic laws for the iterated logarithm law and the bootstrap, when indexed by VC-type classes of kernels. Major (2006) improves these results, for which we present one version in this section. In Clémençon et al. (2008), authors established moment inequalities based on Rademacher chaos, and similar types of results can be found by Adamczak (2006); Giné et al. (2000); Houdré and Reynaud-Bouret (2003). Refer to the monograph De la Pena and Giné (1999) for comprehensive analysis of U-processes. The most recent results are especially interested in applying U-statistics to bootstrap estimators and when applied *e.g.* to censored data. However, the results are mainly obtained for one-sample statistics based on *i.i.d.* random observations, even if the assumptions can imply very advanced techniques.

We consider U-processes defined as follows, with the previously introduced notations. Let $\mathscr{K}$ a class of kernel functions of order (2), U-processes based on a *i.i.d.* sample $\{X_1,\ldots,X_n\}$ are referred to as the mapping

$$k \in \mathscr{K} \mapsto U_n(k) = \frac{1}{n(n-1)}\sum_{1\leq i\neq j\leq n} k(X_i,X_j) \tag{3.3.8}$$

and similarly to empirical process, the goal is to control the uniform deviations of $\{U_n(k) - \theta(k)\}_{k\in\mathscr{K}}$. The selected results give insight into the control of such random object, depending on the type of class of kernels and on the measurability assumption for the uniform bound. Similarly, we refer to U-processes of degree $(1,1)$, based on the two *i.i.d.* and independent samples $\{X_1, \ldots, X_n\}$ and

$\{Y_1, \ldots, Y_m\}$, indexed by a class of kernels $\mathscr{L}$ the collection $\{U_{n,m}(\ell)\}_{\ell\in\mathscr{L}}$.

We start with a maximal inequality proved by Nolan and Pollard (1987) for degenerate U-processes of degree 2 for general classes of symmetric kernels, later extended to two-sample degenerate U-processes of degree $(1,1)$ by Neumeyer (2004). The results are stated in the articles for euclidean classes of kernels, and we express below for VC-type bounded kernels, using respectively for the results Lemma 16 in Nolan and Pollard (1987) and inequality page 83 in Neumeyer (2004), to simplify the bounds of the original Theorems.

Lemma 42 (Consequence of Theorem 6, Nolan and Pollard (1987))**.** *Let $n \geq 2$ and $X_1, \ldots, X_n$ be i.i.d. random variables drawn from a probability distribution μ on a measurable space $\mathscr{X}$. Let $\mathscr{K}$ be a class of measurable kernels $k : \mathscr{X}^2 \to \mathbb{R}$ such that $\sup_{x,x'\in\mathscr{X}^2} |k(x,x')| \leq D < +\infty$ and $\int_{\mathscr{X}^2} k^2(x,x')\mu(dx)\mu(dx') \leq \sigma^2 \leq D^2$, that defines a degenerate one-sample U-process of degree 2, based on the X_i's: $\{U_n(k)\ k \in \mathscr{K}\}$. Suppose in addition that the class $\mathscr{K}$ is of* VC*-type with parameters $(A,\mathscr{V})$. There exists a constant $C > 0$, such that:*

$$\mathbb{E}\left[\sup_{k\in\mathscr{K}} |U_n(k)|\right] \leq \frac{2\sigma C}{n-1}\left(\frac{1}{4} + \mathscr{V}\log(A)\right) . \tag{3.3.9}$$

Lemma 43 (Consequence of Lemma 2.4, Neumeyer (2004))**.** *Let $(n,\ m) \in \mathbb{N}^*$. Consider two independent i.i.d. random samples $X_1,\ldots,X_n$ and $Y_1,\ldots,Y_m$ respectively drawn from the probability distributions μ and ν on the measurable spaces $\mathscr{X}$ and $\mathscr{Y}$. Let $\mathscr{L}$ be a class of degenerate non-symmetrical kernels $\ell : \mathscr{X}\times\mathscr{Y} \to \mathbb{R}$ such that $\sup_{(x,y)\in\mathscr{X}\times\mathscr{Y}} |\ell(x,y)| \leq L < +\infty$ and $\int_{\mathscr{X}\times\mathscr{Y}} \ell^2(x,y)\mu(dx)\nu(dy) \leq \sigma^2 \leq L^2$, that defines a degenerate two-sample U-process of degree $(1,1)$, based on the X_i, Y_j's: $\{U_{n,m}(\ell),\ \ell \in \mathscr{L}\}$. Suppose in addition that the class $\mathscr{L}$ is of* VC*-type with parameters $(A,\mathscr{V})$. There exists a constant $C > 0$, such that:*

$$\mathbb{E}\left[\sup_{\ell\in\mathscr{L}} |U_{n,m}(\ell)|\right] \leq \frac{2\sigma C}{\sqrt{nm}}\left(\frac{1}{4} + \mathscr{V}\log(A)\right) . \tag{3.3.10}$$

In Major (2006) (see Theorem 2 therein), a concentration bound for one-sample degenerate U-processes of arbitrary degree indexed by L_2-dense classes of non-symmetric kernels is established. The lemma below is a formulation of the latter in the specific case of degenerate U-processes of degree 2 indexed by VC-type bounded classes of non-symmetric kernels.

Lemma 44 (Theorem 2, Major (2006))**.** *Suppose the conditions of Lemma 42 fulfilled. Then, there exist constants $C_1 > 0$, $C_2 \geq 1$ and $C_3 \geq 0$ depending on $(A,\mathscr{V})$ such that:*

$$\mathbb{P}\left\{\sup_{k\in\mathscr{K}} |U_n(k)| \geq t\right\} \leq C_2 \exp\left\{-\frac{C_3(n-1)t}{\sigma}\right\} , \tag{3.3.11}$$

as soon as $C_1\log(2D/\sigma) \leq (n-1)t/\sigma \leq n\sigma^2/D^2$.

Part II

Contributions Related to Rank Processes

4 | A Concentration Inequality for Two-sample U-processes

Abstract. This chapter details a new uniform concentration bound obtained for two-sample degenerate U-processes as a preliminary contribution. It is the key to the analysis of R-processes pursued in the following chapters, particularly regarding the nonasymptotic analysis of the remaining term of its decomposition. This new result combines methods related to symmetrization, chaining, and complexity control of VC-type bounded classes of kernels. Therefore, we extend a version of Major (2006)'s results when based on two independent and *i.i.d.* samples drawn from different distributions, as recalled in Chapter 3.

Contents

4.1 Introduction

This chapter studies the supremum of classes of degenerate U-statistics based on two independent samples. Consider their sizes $(n,\, m) \in \mathbb{N}^{*2}$, such that the two independent *i.i.d.* random samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are respectively drawn from probability distributions μ and ν, and valued on the measurable spaces $\mathscr{X}$ and $\mathscr{Y}$. Let $\ell : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$ be a square integrable function *w.r.t.* $\mu \otimes \nu$. The two-sample U-statistic of degree $(1,1)$, with kernel function ℓ and based on the X_i's and the Y_j's is defined by

$$U_{n,m}(\ell) = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \ell(X_i, Y_j) \; . \tag{4.1.1}$$

U-processes refer to collections of U-statistics when indexed by *infinite* classes of kernels. We consider such a class $\mathscr{L}$ composed of kernels $\ell : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$ of controlled complexity thanks to uniform entropic considerations. The purpose is to provide a concentration bound of the deviations of $\sup_{\ell \in \mathscr{L}} |U_{n,m}(\ell)|$. The broad problem of controlling the uniform deviations of U-processes is studied in the literature. However, the results are mainly based on *i.i.d.* random observations while allowing for generalized and various forms of such statistics. We refer, for instance, to Adamczak (2006); Clémençon et al. (2008); Giné et al. (2000); Houdré and Reynaud-Bouret (2003); Major (2006) and the monograph De la Pena and Giné (1999) for fundamental contributions.

In fact, the key to analyze generalized U-statistics (of high degree) relies on Hoeffding's decomposition, as detailed in Theorem 37 (Chap. 3). It allows for the recursive study of U-statistics on their degree, composed of uncorrelated terms with a variance of decreasing order. Nevertheless, the last term needs particular attention as it is of the same order as the initial statistic but known to be degenerate. In this chapter, $U_{n,m}$ is supposed to be degenerate *i.e.* centered when integrating *w.r.t.* the measure μ and ν. Refer to Chapter 3, Section 3.3 for the required properties. The main result of this chapter is the fundamental tool for decomposing linear two-sample R-processes with a uniform control on the remainder process, as will be detailed in the next chapter. We first state the main assumptions and theorem, then the related proofs and additional results are detailed in the following.

4.2 Main result

Consider two independent *i.i.d.* random samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ respectively drawn from the probability distributions μ and ν on the measurable spaces $\mathscr{X}$ and $\mathscr{Y}$. Let $\mathscr{L}$ be a class of degenerate non-symmetrical kernels $\ell : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$ for which the following assumptions are considered.

***Assumption* 1.** *Let $\mathscr{L}$ a class of measurable functions ℓ, and let μ, ν two measures defined on $\mathscr{X}$, $\mathscr{Y}$, such that for all $\ell \in \mathscr{L}$, $\sup_{(x,y) \in \mathscr{X} \times \mathscr{Y}} |\ell(x,y)| \leq L < +\infty$ and $\int_{\mathscr{X} \times \mathscr{Y}} \ell^2(x,y) \mu(dx) \nu(dy) \leq \sigma^2 \leq L^2$.*

***Assumption* 2.** *The class $\mathscr{L}$ is of* VC*-type with parameters $(A, \mathscr{V}) \in (0,\, +\infty)^2$ and of finite constant envelope.*

The main result is stated below and we refer to Chapter 3 for details on the assumptions.

Theorem 45 (Lemma 16, Clémençon et al. (2021)). *Let $(n,\, m) \in \mathbb{N}^*$. Consider two independent i.i.d. random samples $X_1,\ldots,X_n$ and $Y_1,\ldots,Y_m$ respectively drawn from the probability distributions μ and ν on the measurable spaces $\mathscr{X}$ and $\mathscr{Y}$. Let $\mathscr{L}$ be a class of degenerate non-symmetrical kernels $\ell : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$ such that Assumptions 1 and 2 are fulfilled. Then, for all $t > 0$, there exists a universal constant $K > 2$ such that the U-process of degree $(1,1)$ $\{U_{n,m}(\ell),\ \ell \in \mathscr{L}\}$ based on the X_i, Y_j's, satisfies:*

$$\mathbb{P}\left\{\sup_{\ell\in\mathscr{L}} |U_{n,m}(\ell)| \geq t\right\} \leq K 2^{\mathscr{V}} (A/L)^{2\mathscr{V}} e^{4/L^2} \exp\left\{-\frac{nmt^2}{ML^2}\right\}, \tag{4.2.1}$$

for all $nmt^2 > \max(8^4 \log(2) L^2 \mathscr{V}, (\log(2) L^2 \mathscr{V}/2)^{1+\delta})$, $\delta \in (1,2)$ constant and $M = 16^3/2$.

Its proof is given in Section 4.3 and is inspired from that of Lemma 2.14.9 in van der Vaart and Wellner (1996) and of Lemma 3.2 in van de Geer (2000) for empirical processes, and from Lemma 2.4 in Neumeyer (2004) which gives a version in expectation applicable to degenerate two-sample U-processes of arbitrary degree indexed by L_p-dense classes of kernels.

4.3 Proofs

We shall prove an exponential bound of Hoeffding's type for the uniformly bounded two-sample degenerate U-process $\{U_{n,m}(\ell) :\ \ell \in \mathscr{L}\}$, where

$$U_{n,m}(\ell) = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \ell(X_i, Y_j)\ . \tag{4.3.1}$$

In order to apply standard symmetrization arguments, see *e.g.* section 2.3 in van der Vaart and Wellner (1996), consider independent Rademacher variables $\varepsilon_1,\ldots,\varepsilon_n$ and $\eta_1,\ldots,\eta_m$ and define

$$T_{n,m}(\ell) = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \varepsilon_i \eta_j \ell(X_i, Y_j)\ , \tag{4.3.2}$$

for all ℓ in $\mathscr{L}$. We start by proving the following lemmas, involved in the argument.

Lemma 46. *Let P and Q be probability distributions on measurable spaces $\mathscr{X}$ and $\mathscr{Y}$ respectively. Consider the degenerate two-sample U-statistic of degree $(1,1)$ (4.3.1) with a bounded kernel $\ell : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$ based on the independent i.i.d. random samples $X_1,\ \ldots,\ X_n$ and $Y_1,\ \ldots,\ Y_m$, drawn from P and Q respectively. Let two sequences of i.i.d. Rademacher variables $\varepsilon_1,\ldots,\varepsilon_n$ and $\eta_1,\ldots,\eta_m$, independent of the X_i's and Y_j's, such that the randomized process (4.3.2) is defined. Then, for any increasing and convex function $\Phi : \mathbb{R} \to \mathbb{R}$, we have:*

$$\mathbb{E}\left[\Phi\left(\sup_{\ell\in\mathscr{L}} |U_{n,m}(\ell)|\right)\right] \leq \mathbb{E}\left[\Phi\left(4 \sup_{\ell\in\mathscr{L}} |T_{n,m}(\ell)|\right)\right], \tag{4.3.3}$$

and

$$\mathbb{E}\left[\Phi\left(\sup_{\ell\in\mathscr{L}} U_{n,m}(\ell)\right)\right] \leq \mathbb{E}\left[\Phi\left(4 \sup_{\ell\in\mathscr{L}} T_{n,m}(\ell)\right)\right], \tag{4.3.4}$$

assuming that the suprema are measurable and that the expectations exist.

PROOF. We prove the first inequality, the proof of the second one being similar. Using the independence of the two samples, Fubini's theorem and the degeneracy property, one gets that

$$\begin{aligned}
& \mathbb{E}\left[\Phi\left(\sup_{\ell\in\mathscr{L}}|U_{n,m}(\ell)|\right)\right] \\
= \;& \mathbb{E}\left[\mathbb{E}\left[\Phi\left(\sup_{\ell\in\mathscr{L}}\left|\frac{1}{nm}\sum_{i=1}^{n}\left(\sum_{j=1}^{m}\ell(X_i,Y_j)\right)\right|\right) \mid Y_1,\ldots,Y_m\right]\right] \\
\leq \;& \mathbb{E}\left[\Phi\left(2\sup_{\ell\in\mathscr{L}}\left|\frac{1}{nm}\sum_{i=1}^{n}\varepsilon_i\left(\sum_{j=1}^{m}\ell(X_i,Y_j)\right)\right|\right)\right] \\
= \;& \mathbb{E}\left[\mathbb{E}\left[\Phi\left(2\sup_{\ell\in\mathscr{L}}\left|\frac{1}{nm}\sum_{j=1}^{m}\left(\sum_{i=1}^{n}\varepsilon_i\ell(X_i,Y_j)\right)\right|\right) \mid (X_1,\varepsilon_1),\ldots,(X_n,\varepsilon_n)\right]\right] \\
\leq \;& \mathbb{E}\left[\Phi\left(4\sup_{\ell\in\mathscr{L}}\left|\frac{1}{nm}\sum_{j=1}^{m}\eta_j\left(\sum_{i=1}^{n}\varepsilon_i\ell(X_i,Y_j)\right)\right|\right)\right] \\
= \;& \mathbb{E}\left[\Phi\left(4\sup_{\ell\in\mathscr{L}}|T_{n,m}(\ell)|\right)\right]
\end{aligned}$$

by applying Lemma 3.5.2 of De la Pena and Giné (1999) twice. Incidentally, notice that we can also show that

$$\mathbb{E}\left[\Phi\left(\frac{1}{4}\sup_{\ell\in\mathscr{L}}|T_{n,m}(\ell)|\right)\right] \leq \mathbb{E}\left[\Phi\left(\sup_{\ell\in\mathscr{L}}|U_{n,m}(\ell)|\right)\right].$$

by applying twice the reverse inequality in Lemma 3.5.2 of De la Pena and Giné (1999). □

Next, we prove an exponential bound of Hoeffding's type for degenerate two-sample U-statistics with bounded kernels.

Lemma 47. *Let P and Q be probability distributions on measurable spaces $\mathscr{X}$ and $\mathscr{Y}$ respectively. Consider the degenerate two-sample U-statistic of degree $(1,1)$ (4.3.1) with a bounded kernel $\ell : \mathscr{X}\times\mathscr{Y}\to\mathbb{R}$ based on the independent i.i.d. random samples $X_1,\ldots,X_n$ and $Y_1,\ldots,Y_m$, drawn from P and Q respectively. For all $t>0$, we then have:*

$$\mathbb{P}\{U_{n,m}(\ell)\geq t\} \leq e^{-nmt^2/(32c_\ell^2)}, \tag{4.3.5}$$

where $c_\ell = \sup_{(x,y)\in\mathscr{X}\times\mathscr{Y}}|\ell(x,y)| < +\infty$.

PROOF. Let $t>0$. The proof is based on Chernoff's method. For all $\lambda>0$, we have

$$\begin{aligned}
\mathbb{P}\{U_{n,m}(\ell)\geq t\} &\leq \exp\left(-\lambda t + \log\left(\mathbb{E}[\exp(\lambda U_{n,m}(\ell))]\right)\right) \\
&\leq \exp\left(-\lambda t + \log\left(\mathbb{E}[\exp(4\lambda T_{n,m}(\ell))]\right)\right), \qquad (4.3.6)
\end{aligned}$$

using (4.3.4) with $\Phi(t)=\exp(\lambda t)$. Observe next that we almost-surely

$$\begin{aligned}
&\mathbb{E}[\exp(4\lambda T_{n,m}(\ell)) \mid X_1,\ldots,X_n,Y_1,\ldots,Y_m] = \\
&\qquad \prod_{i=1}^{n}\prod_{j=1}^{m}\frac{e^{4\lambda\ell(X_i,Y_j)/(nm)}+e^{-4\lambda\ell(X_i,Y_j)/(nm)}}{2} \\
&\qquad\qquad \leq \prod_{i=1}^{n}\prod_{j=1}^{m} e^{8\lambda^2\ell^2(X_i,Y_j)/(nm)^2} \leq e^{8\lambda^2 c_\ell^2/(nm)},
\end{aligned}$$

using the fact that $(e^u + e^{-u})/2 \leq e^{u^2/2}$ for all $u \in \mathbb{R}$. Integrating the bound over the X_i's and Y_j's and plugging it next into (4.3.6) yields the desired bound when choosing $\lambda = nmt/(16c_\ell^2)$.
□

Finally, we prove the tail probability version of Lemma 46 stated below.

Lemma 48. *Let P and Q be probability distributions on measurable spaces $\mathscr{X}$ and $\mathscr{Y}$ respectively. Consider the degenerate two-sample U-statistic of degree $(1,1)$ (4.3.1) with a bounded kernel $\ell : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$ based on the independent i.i.d. random samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, drawn from P and Q respectively. Let two sequences of i.i.d. Rademacher variables $\varepsilon_1, \ldots, \varepsilon_n$ and $\eta_1, \ldots, \eta_m$, independent of the $X$is and Y js, such that the randomized process (4.3.2) is defined. Then we have for all $t > 0$,*

$$\mathbb{P}\left\{\sup_{\ell\in\mathscr{L}} |U_{n,m}(\ell)| \geq 16t\right\} \leq 16\mathbb{P}\left\{\sup_{\ell\in\mathscr{L}} |T_{n,m}(\ell)| \geq t\right\}, \tag{4.3.7}$$

assuming that the suprema are measurable and that the expectations exist.

PROOF. This lemma, bounding the tail probability of $\sup_{\ell\in\mathscr{L}} |U_{n,m}(\ell)|$ to that of $\sup_{\ell\in\mathscr{L}} |T_{n,m}(\ell)|$, generalizes Lemma 2.7 in Giné and Zinn (2004) and Lemma 3.1 in Talagrand (1994) to degenerate two-sample U-processes. It is proved by applying twice a version of the latter result for independent but non necessarily identically distributed random variables. Indeed, we have: $\forall t > 0$,

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{\ell\in\mathscr{L}} |U_{n,m}(\ell)| \geq 16t\right\} \\
&\qquad = \mathbb{E}\left[\mathbb{P}\left\{\sup_{\ell\in\mathscr{L}} \left|\frac{1}{n}\sum_{i=1}^n \left\{\frac{1}{m}\sum_{j=1}^m \ell(X_i,Y_j)\right\}\right| \geq 16t \mid Y_1, \ldots, Y_m\right\}\right] \\
&\qquad \leq 4\mathbb{E}\left[\mathbb{P}\left\{\sup_{\ell\in\mathscr{L}} \left|\frac{1}{n}\sum_{i=1}^n \left\{\frac{1}{m}\sum_{j=1}^m \varepsilon_i\ell(X_i,Y_j)\right\}\right| \geq 4t \mid Y_1, \ldots, Y_m\right\}\right] \\
&\qquad = 4\mathbb{E}\left[\mathbb{P}\left\{\sup_{\ell\in\mathscr{L}} \left|\frac{1}{m}\sum_{j=1}^m \left\{\frac{1}{n}\sum_{i=1}^n \varepsilon_i\ell(X_i,Y_j)\right\}\right| \geq 4t \mid (X_1,\varepsilon_1) \ldots, (X_n, \varepsilon_n)\right\}\right] \\
&\qquad \leq 16\mathbb{P}\left\{\sup_{\ell\in\mathscr{L}} |T_{n,m}(\ell)| \geq t\right\}.
\end{aligned}$$

□

The proof relies on the chaining method applied to the process $U_{n,m}(\ell)$ indexed by the class of kernels $\mathscr{L}$, see *e.g.* the argument used to establish Lemma 2.14.9 in van der Vaart and Wellner (1996). Define the random semi-metric on $\mathscr{L}$ by

$$d_{nm}^2(\ell_1,\ell_2) = \frac{1}{nm}\sum_{i\leq n}\sum_{j\leq m} (\ell_1(X_i,Y_j) - \ell_2(X_i,Y_j))^2 \tag{4.3.8}$$

for all kernels ℓ_1 and ℓ_2 in $\mathscr{L}$. For all $q \in \mathbb{N}^*$, consider a number $k_q \leq (A/\varepsilon_q)^\nu$ of L_2-balls with radius $\varepsilon_q \leq L \leq 1$ and centers $\ell_{q,k}$, $1 \leq k \leq k_q$, *w.r.t.* the (random) probability measure $(1/nm)\sum_{i\leq n}\sum_{j\leq m}\delta_{(X_i,Y_j)}$ covering the class $\mathscr{L}$. Assume that the sequence ε_q is decreasing as q increases, so that k_q is increasing. Let $\ell \in \mathscr{L}$, $q \geq 1$ and $\tilde{\ell}_q$ be the center of a ball s.t. $d_{nm}(\ell,\tilde{\ell}_q) \leq \varepsilon_q$. Fixing $q_0 \leq q$ in $\mathbb{N}^*$, the following decomposition holds

$$U_{n,m}(\ell) = (U_{n,m}(\ell) - U_{n,m}(\tilde{\ell}_q)) + U_{n,m}(\tilde{\ell}_{q_0}) + \sum_{\omega=q_0+1}^{q} \left(U_{n,m}(\tilde{\ell}_\omega) - U_{n,m}(\tilde{\ell}_{\omega-1})\right).$$

Observe that, for all ℓ in $\mathscr{L}$, we almost-surely have

$$|U_{n,m}(\ell) - U_{n,m}(\tilde{\ell}_q)| \leq d_{nm}(\ell, \tilde{\ell}_q) \leq \varepsilon_q \ .$$

The triangular inequality yields

$$\|U_{n,m}(\ell)\|_{\mathscr{L}} \leq \varepsilon_q + \max_{1\leq k\leq k_{q_0}} |U_{n,m}(\ell_{q_0,k})| + \sum_{\omega=q_0+1}^{q} \|U_{n,m}(\tilde{\ell}_\omega) - U_{n,m}(\tilde{\ell}_{\omega-1})\|_{\mathscr{L}} \ ,$$

where we used the notation $\|V\|_{\mathscr{L}} = \sup_{\ell\in\mathscr{L}} |V(\ell)|$ for any real-valued stochastic process V indexed by $\mathscr{L}$. Considering $\eta_\omega > 0$ and $\beta > 0$ constants such that $\sum_{\omega=q_0+1}^{q} \eta_\omega + \beta \leq 1$, we have for any $t > \varepsilon_q$:

$$\begin{aligned}\mathbb{P}\{\|U_{n,m}(\ell)\|_{\mathscr{L}} \geq 16t\} &\leq \sum_{k=1}^{k_{q_0}} \mathbb{P}\left\{|U_{n,m}(\ell_{q_0,k}))| \geq 16t\beta\right\} \\ &\quad + 16 \sum_{\omega=q_0+1}^{q} k_\omega^2 \mathbb{E}\left[\sup_{\ell\in\mathscr{L}} \mathbb{P}\left\{|T_{n,m}(\tilde{\ell}_\omega - \tilde{\ell}_{\omega-1})| \geq t\eta_\omega \mid X_1, \ldots, X_n, Y_1, \ldots, Y_m\right\}\right], \quad (4.3.9)\end{aligned}$$

using the union bound, Lemma 48 and observing that the suprema corresponding to the terms of the series are actually maxima taken over at most $k_\omega k_{\omega-1} \leq k_\omega^2$ elements. Lemma 47 permits to bound the first term on the right hand side of (4.3.9):

$$\sum_{k=1}^{k_{q_0}} \mathbb{P}\left\{|U_{n,m}(\ell_{q_0,k}))| \geq 16t\beta\right\} \leq 2k_{q_0} \exp\left\{-\frac{8nm(t\beta)^2}{L^2}\right\} . \quad (4.3.10)$$

Concerning the second term, notice that

$$d_{nm}(\tilde{\ell}_\omega, \tilde{\ell}_{\omega-1}) \leq d_{nm}(\ell, \tilde{\ell}_{\omega-1}) + d_{nm}(\tilde{\ell}_\omega, \ell) \leq 2\varepsilon_{\omega-1} \ . \quad (4.3.11)$$

Re-using the start of the argument proving Lemma 47, we have: $\forall\lambda > 0$,

$$\begin{aligned}&\mathbb{P}\left\{T_{n,m}(\tilde{\ell}_\omega - \tilde{\ell}_{\omega-1}) \geq t\eta_\omega \mid X_1, \ldots, X_n, Y_1, \ldots, Y_m\right\} \\ &\qquad \leq \exp\left(-\lambda t\eta_\omega + \mathbb{E}\left[\exp(\lambda T_{n,m}(\tilde{\ell}_\omega - \tilde{\ell}_{\omega-1})) \mid X_1, \ldots, X_n, Y_1, \ldots, Y_m\right]\right)\end{aligned}$$

with probability one. Like in Lemma 47's proof, we almost-surely have

$$\begin{aligned}&\mathbb{E}[\exp(\lambda T_{n,m}(\tilde{\ell}_\omega - \tilde{\ell}_{\omega-1})) \mid X_1, \ldots, X_n, Y_1, \ldots, Y_m] \leq \\ &\qquad \prod_{i=1}^{n}\prod_{j=1}^{m} e^{\lambda^2(\tilde{\ell}_\omega - \tilde{\ell}_{\omega-1})^2(X_i,Y_j)/2(nm)^2} \leq e^{2\lambda^2\varepsilon_{\omega-1}^2/(nm)} .\end{aligned}$$

Combining the two bounds above with the union bound, it holds with probability one

$$\mathbb{P}\left\{\left|T_{n,m}(\tilde{\ell}_\omega - \tilde{\ell}_{\omega-1})\right| \geq t\eta_\omega \mid X_1, \ldots, X_n, Y_1, \ldots, Y_m\right\} \leq 2\exp\left\{-\frac{nm(t\eta_\omega)^2}{8\varepsilon_{\omega-1}^2}\right\} . \quad (4.3.12)$$

From (4.3.9), (4.3.10) and (4.3.12), we deduce that

$$\mathbb{P}\{\|U_{n,m}(\ell)\|_{\mathscr{L}} \geq 16t\}$$
$$\leq 2k_{q_0}\exp\left\{-\frac{8nm(t\beta)^2}{L^2}\right\}+32\sum_{\omega=q_0+1}^{q}k_\omega^2\exp\left\{-\frac{nm(t\eta_\omega)^2}{8\varepsilon_{\omega-1}^2}\right\}$$
$$\leq 2A^{\mathscr{V}}\varepsilon_{q_0}^{-\mathscr{V}}\exp\left\{-\frac{8nm(t\beta)^2}{L^2}\right\}+32A^{2\mathscr{V}}\sum_{\omega=q_0+1}^{q}\varepsilon_\omega^{-2\mathscr{V}}\exp\left\{-\frac{nm(t\eta_\omega)^2}{8\varepsilon_{\omega-1}^2}\right\}. \quad (4.3.13)$$

Following Lemma 3.2 in van de Geer (2000) and choosing $\varepsilon_\omega = 2^{-\omega}L$, $\eta_\omega = 2^{-\omega}\sqrt{\omega}/8$, so that $\eta_{\omega+1}/\varepsilon_\omega = (1/16L)\sqrt{\omega+1}$, we have

$$\varepsilon_\omega^{-2\mathscr{V}}\exp\left\{-\frac{nm(t\eta_\omega)^2}{8\varepsilon_{\omega-1}^2}\right\} = L^{-2\mathscr{V}}\exp\left\{-(-2\mathscr{V}\log(2)+\frac{nmt^2}{4\times 8^3L^2})\omega\right\} \quad (4.3.14)$$

If $nmt^2 > 8^4\log(2)L^2\mathscr{V}$, the terms of the series are decreasing *w.r.t.* ω and we upperbound by $K_1L^{-2\mathscr{V}}\exp\{-nmt^2\omega/(4\times 8^3L^2)\}$. Problem 2.14.3 in van der Vaart and Wellner (1996) applies for $\omega\in\{q_0+1,\ldots,q\}$ with $\psi(\omega) = nmt^2\omega/(4\times 8^3L^2)$

$$\sum_{\omega=q_0+1}^{q}\varepsilon_\omega^{-2\mathscr{V}}\exp\left\{-\frac{nm(t\eta_\omega)^2}{8\varepsilon_{\omega-1}^2}\right\} \leq K_1L^{-2\mathscr{V}}\psi'(q_0)^{-1}\exp\{-\psi(q_0)\}$$
$$\leq K_2L^{-2(\mathscr{V}-1)}\exp\left\{-\frac{nmt^2}{4\times 8^3L^2}q_0\right\} \quad (4.3.15)$$

K_1, $K_2 > 0$ constants and $nmt^2 \geq 1$. For $\alpha > 0$ large, setting $q_0 = 2+\lfloor(nmt^2)^{1/(\alpha-1)}\rfloor$ yields to the upperbound $K_2L^{-2(\mathscr{V}-1)}\exp\{-3nmt^2/(4\times 8^3L^2)\}$. For the first tail probability, by setting $\beta = 1/2-1/(2nmt^2)$ we obtain an upperbound of similar form

$$A^{\mathscr{V}}\varepsilon_{q_0}^{-\mathscr{V}}\exp\left\{-\frac{8nm(t\beta)^2}{L^2}\right\}$$
$$\leq (A/L)^{\mathscr{V}}\exp\left\{\mathscr{V}\log(2)(2+(nmt^2)^{1/(\alpha-1)})-\frac{2nmt^2}{L^2}(1-1/(nmt^2))^2\right\}$$
$$\leq (2A/L)^{\mathscr{V}}e^{4/L^2}\exp\left\{\mathscr{V}\log(2)(nmt^2)^{1/(\alpha-1)}-\frac{2nmt^2}{L^2}\right\}$$
$$\leq (2A/L)^{\mathscr{V}}e^{4/L^2}\exp\left\{-\frac{2nmt^2}{L^2}\right\},$$

as soon as $nmt^2 > (\log(2)L^2\mathscr{V}/2)^{1+\delta}$, $\delta = 1/(\alpha-2)\in(0,1)$ for large α. Gathering both upperbounds, Eq. (4.3.13) yields

$$\mathbb{P}\{\|U_{n,m}(\ell)\|_{\mathscr{L}} \geq t\} \leq K2^{\mathscr{V}+1}(A/L)^{2\mathscr{V}}e^{4/L^2}\exp\left\{-\frac{3nmt^2}{4\times 8^3L^2}\right\}, \quad (4.3.16)$$

for all $nmt^2 > \max(1, 8^4\log(2)L^2\mathscr{V}, (\log(2)L^2\mathscr{V}/2)^{1+\delta})$, and $K \geq 1+16K_2e^{-4}$ constant. Checking lastly that, for all $q \geq 1$

$$8\sum_{\omega=q_0+1}^{q}\eta_\omega \leq 8\sum_{\omega=1}^{q}\eta_\omega \leq 1+\int_1^\infty 2^{-x}\sqrt{x}dx \leq 1+(\pi/\log(2))^{1/2} \leq 4, \quad (4.3.17)$$

so that $\sum_{\omega=q_0+1}^{q}\eta_\omega+\beta \leq 1$ as needed.

Permanence properties for classes of functions. The lemma stated below permits to control the complexity of the classes of kernels/functions involved in the Hoeffding decompositions of a two-sample U-process of degree $(1,1)$.

Lemma 49. *Let X and Y be two independent random variables, valued in $\mathscr{X}$ and $\mathscr{Y}$ respectively, with probability distributions μ and ν. Consider $\mathscr{L}$ a* VC*-type bounded class of kernels $\ell : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$ with parameters $(A, \mathscr{V})$ and constant envelope $L_{\mathscr{L}} > 0$. Then, the sets of functions $\{x \in \mathscr{X} \mapsto \mathbb{E}[\ell(x, Y)] : \ell \in \mathscr{L}\}$, $\{y \in \mathscr{Y} \mapsto \mathbb{E}[\ell(X, y)] : \ell \in \mathscr{L}\}$, $\{\ell(x,y) - \mathbb{E}[\ell(X, y)] - \mathbb{E}[\ell(x, Y)] : \ell \in \mathscr{L}\}$ are also* VC*-type bounded classes.*

Remark 2. *Notice that Lemma 49 remains true if one considers two indendepent r.v. X, X' valued in $\mathscr{X} = \mathscr{Y}$ and $\mu = \nu$. Hence, this will be similarly used for one-sample U-processes of degree 2, as introduced in Section 3.3 Chapter 3 therein.*

PROOF. Consider first the uniformly bounded class $\mathscr{L}_1$ composed of functions $x \in \mathscr{X} \mapsto \mathbb{E}[\ell(x, Y)]$ with $\ell \in \mathscr{L}$. Let $\varepsilon > 0$ and P be any probability measure on $\mathscr{X}$. Define the probability measure $P_\nu(dx,dy) = P(dx)\nu(dy)$ on $\mathscr{X} \times \mathscr{Y}$ and consider a ε-covering of the class $\mathscr{L}$ with centers $\ell_1, \ldots, \ell_K$ *w.r.t.* the metric $L_2(P_\nu)$, $K \geq 1$. For all $\ell \in \mathscr{L}$, there exists $k \leq K$ such that:

$$\int_{x \in \mathscr{X}} \int_{y \in \mathscr{Y}} (\ell(x, y) - \ell_k(x, y))^2 P_\nu(dx,dy) \leq \varepsilon^2 .$$

By virtue of Jensen's inequality, we have

$$\begin{aligned}\int_{\mathscr{X}} (\mathbb{E}[\ell(x, Y)] - \mathbb{E}[\ell_k(x, Y)])^2 P(dx) &\leq \int_{\mathscr{X}} \mathbb{E}[(\ell(x, Y) - \ell_k(x, Y))^2] P(dx) \\ &= \int_{\mathscr{X}} \int_{\mathscr{Y}} (\ell(x, y) - \ell_k(x, y))^2 \nu(dy) P(dx) \leq \varepsilon^2 .\end{aligned}$$

Hence, one gets a ε-covering of the class $\mathscr{L}_1$ with balls of centers $\{\mathbb{E}[\ell_k(\cdot, Y)] : k = 1, \ldots, K\}$ in $L_2(P)$. This proves that

$$N(\mathscr{L}_1, L_2(P), \varepsilon L_{\mathscr{L}}) \leq N(\mathscr{L}, L_2(P_\nu), \varepsilon L_{\mathscr{L}}).$$

As a similar reasoning can be applied to the two other classes of functions, one then gets the desired result. □

5 | Concentration Inequalities for Two-sample R-processes

Abstract. This chapter studies a new class of *two-sample linear rank statistics* for the multivariate and nonparametric framework. We first motivate this approach by showing how it encompasses and summarizes empirical criteria of classical two-sample problems, as detailed in Chapter 2. In particular, we highlight how it is intimately related to the ROC analysis. Briefly, the ROC curve is the gold standard for measuring the performance of a test/scoring statistic regarding its capacity to discriminate between two statistical populations in a wide variety of applications, ranging from anomaly detection in signal processing to information retrieval, through medical diagnosis. Most practical performance measures used in scoring/ranking applications such as the AUC, the local AUC, the p-norm push, the DCG and others, can be viewed as summaries of the ROC curve. Then, concentration inequalities for collections of such random variables, referred to as *two-sample rank processes*, are proved, when indexed by VC classes of scoring functions. Based on these nonasymptotic bounds, the generalization capacity of empirical maximizers of a wide class of ranking performance criteria is next investigated from a theoretical perspective. The numerical experiments are gathered in Chapter 7.

Contents

5.1 Introduction

In the context of ranking, a variety of performance measures can be considered. In the simplest framework of bipartite ranking, where two independent *i.i.d.* samples $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ defined on the same probability space $(\Omega, \mathscr{F}, \mathbb{P})$, valued in the same space $\mathscr{Z}$, say $\mathbb{R}^d$ with $d \geq 1$ for instance, and drawn from probability distributions G and H respectively (referred to as the 'positive distribution' and the 'negative distribution' respectively), the goal pursued is to learn a preorder on $\mathscr{Z}$ defined through a scoring function $s : \mathscr{Z} \to \mathbb{R}$ (which transports the natural order on the real line onto the feature space $\mathscr{Z}$) such that, for any random observation $\mathbf{Z} \in \mathscr{Z}$ sampled from a distribution that is equal either to the 'positive distribution' or to the 'negative one', the larger the score $s(z)$, the likelier it is drawn from the 'positive distribution' G. Though easy to formulate, this simple framework encompasses many practical problems from the design of search engines in Information Retrieval (in this case, for a specific request, G is the distribution of the relevant digitized documents, while H is that of the irrelevant ones) to the elaboration of decision support tools in personalized medicine for instance. In spite of its simplicity there is not one and only one natural scalar criterion for evaluating the performance of a scoring rule $s(z)$, but many possible options. The *Receiving Operator Characteric* curve (the ROC curve in abbreviated form), *i.e.* the PP-plot of the false positive rate *vs* the true positive rate:

$$t \in \mathbb{R} \mapsto \left(\mathbb{P}\{s(\mathbf{Y}) > t\}, \mathbb{P}\{s(\mathbf{X}) > t\}\right),$$

denoting by $\mathbf{X}$ and $\mathbf{Y}$ two generic *r.v.* with distributions G and H respectively, provides an exhaustive description of the performance of any scoring rule candidate s. However, its functional nature renders direct optimization strategies rather complex, see *e.g.* Clémençon and Vayatis (2010). *Empirical risk minimization* methods (ERM) are thus generally based on summaries of the ROC curve, which take the form of empirical risk functionals where the averages involved are no longer taken over *i.i.d.* sequences. The most popular choice is undoubtedly the AUC criterion (AUC standing

for *Area Under the* ROC *Curve*), see Agarwal et al. (2005) or Clémençon et al. (2008) for instance, but when focus is on top-ranked instances, various choices can be considered, *e.g.* the Discounted Cumulative Gain or DCG (see Cossock and Zhang (2006)), the p-norm push (see Rudin (2006)), the local AUC (refer to Clémençon and Vayatis (2007)) or other variants such as those recently introduced in Menon and Williamson (2016). The present chapter starts from the simple observation that most of these summary criteria have a common feature: they belong to the class of *two-sample linear rank statistics*. Such statistics have been extensively studied in the mathematical statistics literature because of their optimality properties in hypothesis testing, see Hájek and Sidák (1967). They are widely used in order to test whether two samples are drawn from the same distribution against the alternative stipulating that the distribution of one of the samples is stochastically larger than the other. For instance, the empirical counterpart of the AUC of a scoring function $s(z)$ corresponds to the popular Mann-Withney-Wilcoxon statistic based on the two (univariate) samples $s(\mathbf{X}_1), \ldots, s(\mathbf{X}_n)$ and $s(\mathbf{Y}_1), \ldots, s(\mathbf{Y}_m)$. Other rank statistics can be considered, corresponding to other ways of measuring how the distribution of the 'positive score' $s(\mathbf{X})$ is (possibly) stochastically larger than that of the 'negative score' $s(\mathbf{Y})$. Now, in the statistical learning view, with the importance of excess risk bounds, the *Empirical Risk Minimization* paradigm must be revisited and new problems, mainly related to the uniform control of the fluctuations of collections of two-sample linear rank statistics, termed rank processes throughout the chapter, and to the measure of the complexity of nonparametric classes of scoring functions, come up. The arguments required to deal with risk functionals based on two-sample linear rank statistics have been sketched in Clémençon and Vayatis (2007) in a very special case.

In the present chapter, we relate two-sample linear rank statistics to performance measures relevant for the ranking problem by showing that the target of ranking algorithms corresponds to optimal ordering rules in this sense and show that the generic structure of two-sample linear rank statistics as an orthogonal decomposition after projection onto the space of sums of *i.i.d.* random variables is the key to all statistical results related to maximizers of such criteria: consistency, rates of convergence or model selection. Notice incidentally that the empirical AUC is also a U-statistic and a decomposition method akin to that considered in this chapter (though much less general) has been used in order to handle this specific dependence structure in Clémençon et al. (2008). In this chapter, concentration properties of two-sample rank processes (*i.e.* collections of two-sample linear rank statistics) are investigated using the linearization technique aforementioned. While interesting in themselves, the concentration inequalities established for this class of stochastic processes, when indexed by Vapnik-Chervonenkis classes (abbreviated with VC-classes) of scoring functions, are next applied to study the generalization capacity of empirical maximizers of a large collection of performance criteria based on two-sample linear rank statistics. Notice finally that a preliminary version of this work is briefly outlined in the conference paper Clémençon and Vayatis (2009a). This chapter presents a much deeper analysis of bipartite ranking via maximization of two-sample linear rank statistics. In particular, it offers a complete and detailed study of the concentration properties of two-sample rank processes (in a slightly different framework, stipulating that two independent *i.i.d.* samples, positive and negative, are observed, rather than classification data), provides model selection results and, from a practical perspective, tackles the issue of smoothing the risk functionals under study here with statistical learning guarantees.

The chapter is organized as follows. In Section 5.2, the main notations are set out, the bipartite ranking problem is formulated as a statistical learning task in a rigorous probabilistic framework and the concept of two-sample linear rank statistic is briefly recalled. It is also explained that, unsurprisingly, natural performance criteria in bipartite ranking are of the form of two-sample (linear) rank statistics. Concentration results for rank processes, are established in Section 5.3. By means of the latter, performance of optimal scoring functions obtained by maximizing two-sample linear rank

statistics are investigated in Section 5.4. Proofs, technical details and additional numerical results are deferred to the Appendix section. Lastly, numerical experiments are gathered in the Chapter 7 for the bipartite ranking application.

5.2 Motivation and preliminaries

This section motivates the chosen formulation of two-sample rank processes by virtue of the bipartite ranking problem and its relation to the ROC analysis. Indeed, having introduced the (univariate) ROC analysis and its relation to bipartite ranking in section 2.2, we show how two-sample linear rank statistics are empirical scalar summaries of classical related criteria, commonly used as bipartite ranking .

5.2.1 Two-sample linear rank statistics

In a generic formulation, by considering two univariate distributions G and H, the ROC curve, defined by $\mathrm{ROC}_{\mathrm{H,G}}$, is a functional criterion that examines to which extent G is stochastically larger than H. A possible parametric definition in $[0,1]^2$ is given by:

$$t \in \mathbb{R} \mapsto (1-H(t),\, 1-G(t))\,,$$

with the convention that possible jumps are connected by line segments, ensuring that the resulting curve is continuous. With this convention, one may then see the ROC curve related to the pair of *d.f.* (H,G) as the graph of a càd-làg (*i.e.* right-continuous and left-limited) non decreasing mapping valued in $[0,1]$, defined by:

$$\alpha \in (0,1) \mapsto 1 - G \circ H^{-1}(1-\alpha)\,,$$

at points α such that $H \circ H^{-1}(1-\alpha) = 1-\alpha$. See Appendix section B.3 for the univariate definitions (B.3.1) and (B.3.2). However, practical decisions are generally made on the basis of the observations of two univariate independent random *i.i.d.* samples $\{X_1,\, \ldots,\, X_n\}$ and $\{Y_1,\, \ldots,\, Y_m\}$, drawn from G and H respectively. Computing the empirical cumulative distribution functions $\widehat{H}_m(t) = (1/m)\sum_{j=1}^{m} \mathbb{I}\{Y_j \leq t\}$ and $\widehat{G}_n(t) = (1/n)\sum_{i=1}^{n} \mathbb{I}\{X_i \leq t\}$ for $t \in \mathbb{R}$, one can plot the empirical ROC curve:

$$\widehat{\mathrm{ROC}} = \mathrm{ROC}_{\widehat{\mathrm{H}}_{\mathrm{m}},\, \widehat{\mathrm{G}}_{\mathrm{n}}}\,. \tag{5.2.1}$$

Observe that the ROC curve (5.2.1) is an increasing broken line connecting $(0,0)$ to $(1,1)$ in the unit square $[0,1]^2$ and is fully determined by the set of ranks occupied by the positive instances within the pooled sample $\{\mathrm{Rank}(X_i) :\ i = 1,\, \ldots,\, n\}$, where:

$$\forall i \in \{1,\, \ldots,\, n\} \quad \mathrm{Rank}(X_i) = N\widehat{F}_N(X_i)\,, \tag{5.2.2}$$

with $\widehat{F}_N(t) = (1/N)\sum_{i=1}^{n} \mathbb{I}\{X_i \leq t\} + (1/N)\sum_{j=1}^{m} \mathbb{I}\{Y_j \leq t\}$ and $N = n+m$. Breakpoints of the piecewise linear curve (5.2.1) necessarily belong to the set of gridpoints

$$\{(j/m,\, i/n) :\ j \in \{1,\, \ldots,\, m-1\} \text{ and } i \in \{1,\, \ldots,\, n-1\}\}\,.$$

Denote by $X_{(i)}$ the order statistics related to the sample $\{X_1,\, \ldots,\, X_n\}$, *i.e.* $\mathrm{Rank}(X_{(n)}) > \cdots > \mathrm{Rank}(X_{(1)})$, and by $Y_{(j)}$ those related to the sample $\{Y_1,\, \ldots,\, Y_m\}$. Consider the càd-làg step function:

$$\alpha \in [0,1] \mapsto \sum_{j=1}^{m} \widehat{\gamma}_j \cdot \mathbb{I}\{\alpha \in [(j-1)/m,\, j/m[\}\,, \tag{5.2.3}$$

where, for all $j \in \{1, \ldots, m\}$, we set:

$$\widehat{\gamma}_j = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{X_i > Y_{(m-j+1)}\} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{\mathrm{Rank}(X_{(n-i+1)}) > \mathrm{Rank}(Y_{(m-j+1)})\}$$
$$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{j \geq N - \mathrm{Rank}(X_{(n-i+1)}) - i + 2\} \, .$$

The ROC curve (5.2.1) is the continuous broken line that connects the jump points of the step curve (5.2.3) and can thus be expressed as a function of the 'positive ranks' *i.e.* the $\mathrm{Rank}(X_i)$'s only. As a consequence, any summary of the empirical ROC curve, is a two-sample rank statistic, that is a measurable function of the 'positive ranks'. In particular, the empirical AUC, *i.e.* the AUC of the empirical ROC curve (5.2.1), also termed the rate of consistent pairs or the *Mann-Whitney statistic*, can be easily shown to coincide, up to an affine transform, with the sum of 'positive ranks', the well-known *rank-sum Wilcoxon statistic* Wilcoxon (1945)

$$\widehat{W}_{n,m} = \sum_{i=1}^{n} \mathrm{Rank}(X_i) \, . \tag{5.2.4}$$

Indeed, we have

$$\widehat{W}_{n,m} = nm\mathrm{AUC}_{\widehat{\mathrm{H}}_\mathrm{m},\widehat{\mathrm{G}}_\mathrm{n}} + \frac{\mathrm{n}(\mathrm{n}+1)}{2} \, . \tag{5.2.5}$$

However, two-sample rank statistics (*i.e.* functions of the $\mathrm{Rank}(X_i)$'s) form a very rich collection of statistics and this is by no means the sole possible choice to summarize the empirical ROC curve.

Definition 50. (TWO-SAMPLE LINEAR RANK STATISTICS) *Let $\phi : [0,1] \to \mathbb{R}$ be a nondecreasing function. The two-sample linear rank statistics with 'score-generating function' $\phi(u)$ based on the random samples $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_m\}$ is given by:*

$$\widehat{W}^{\phi}_{n,m} = \sum_{i=1}^{n} \phi\left(\frac{\mathrm{Rank}(X_i)}{N+1}\right) . \tag{5.2.6}$$

The statistics (8.3.3) defined above are all distribution-free when $H = G$ and are, for this reason, particularly useful to detect differences between the distributions H and G and widely used to perform homogeneity tests in the univariate setup. Tabulating their distribution under the null assumption, they can be used to design unbiased tests at certain levels α in $(0,1)$. The choice of the score-generating function ϕ can be guided by the type of difference between the two distributions (*e.g.* in scale, in location) one possibly expects, and may then lead to locally most powerful testing procedures, capable of detecting 'small' deviations from the homogeneous situation. More generally, depending on the statistical test to perform, one may use particular function ϕ, Figure 5.1 shows classic score-generating functions broadly used for two-sample statistical tests (refer to Hájek (1962)). One may refer to Chapter 9 in Serfling (1980) or to Chapter 13 in van der Vaart (1998) for an account of the (asymptotic) theory of rank statistics.

Alternatively, two-sample linear rank statistics can be used for a very different purpose, as empirical performance measures in bipartite ranking based on two independent multivariate samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$. The analysis of the bipartite ranking problem carried out in Section 5.4, based on the concentration inequalities established in Section 5.3, shows the relevance of evaluating the ranking performance of a scoring rule candidate $s(z)$ by computing a two-sample linear rank statistic based on the univariate samples obtained after scoring $\{s(\mathbf{X}_1), \ldots, s(\mathbf{X}_n)\}$ and $\{s(\mathbf{Y}_1), \ldots, s(\mathbf{Y}_m)\}$ and establishes statistical guarantees for the generalization capacity of scoring rules built by optimizing such an empirical criterion.

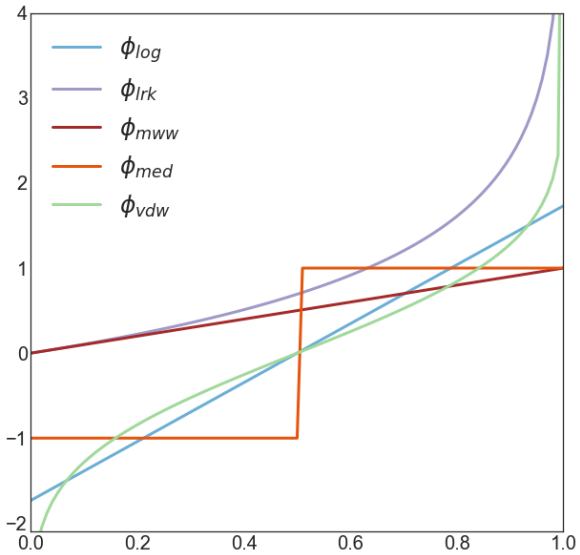


Figure 5.1. Curves of two-sample score-generating functions with the associated statistical test: Logistic test $\phi_{log}(u) = 2\sqrt{3}(u-1/2)$ in blue, Logrank test $\phi_{lrk}(u) = -\log(1-x)$ in purple, Mann- Whitney-Wilcoxon test $\phi_{mww}(u) = u$ in red, Median test $\phi_{med}(u) = \mathrm{sgn}(u-1/2)$ in orange, Van der Waerden test $\phi_{vdw}(u) = \Phi^{-1}(u)$ in green, Φ being the normal quantile function.

5.2.2 Bipartite ranking as maximization of two-sample rank statistics

Going back to the multivariate setup, where H and G are probability distributions on $\mathscr{Z}$, say $\mathscr{Z} = \mathbb{R}^d$ with arbitrary dimension $d \geq 1$, the goal pursued in bipartite ranking can be phrased as that of building a scoring rule $s(z)$ such that the (univariate) distribution G_s of $s(\mathbf{X})$ is 'as stochastically larger as possible' than the distribution H_s of $s(\mathbf{Y})$. Hence, the capacity of a candidate $s(z)$ to discriminate between the positive and negative statistical populations can be evaluated by plotting the ROC curve $\alpha \in (0,1) \mapsto \mathrm{ROC}(\mathrm{s},\alpha) = \mathrm{ROC}_{\mathrm{H_s,G_s}}(\alpha)$: the closer to the left upper corner of the unit square the curve $\mathrm{ROC}(\mathrm{s},.)$, the better the scoring rule s. Therefore, the ROC curve conveys a partial preorder on the set of all scoring functions: for all pairs of scoring functions s_1 and s_2, one says that s_2 is more accurate than s_1 when $\mathrm{ROC}(\mathrm{s}_1,\alpha) \leq \mathrm{ROC}(\mathrm{s}_2,\alpha)$ for all $\alpha \in [0,1]$.

It follows from a standard Neyman-Pearson argument that the most accurate scoring rules are increasing transforms of the likelihood ratio $\Psi(z) = dG/dH(z)$. Precisely, it is shown in Clémençon and Vayatis (2009b) (see Proposition 2 therein) that the optimal scoring rules are the elements of the set:

$$\mathscr{S}^* = \left\{ s \in \mathscr{S} \ \ s.t. \text{ for all } z,\, z' \text{ in } \mathbb{R}^d : \ \Psi(z) < \Psi(z') \Rightarrow s^*(z) < s^*(z') \right\}. \tag{5.2.7}$$

We denote by $\mathrm{ROC}^*(.) = \mathrm{ROC}(\Psi,\, .)$ and recall incidentally that this optimal curve is non-decreasing and concave and thus always above the main diagonal of the unit square. The bipartite ranking task can be reformulated in a more quantitative manner: the objective pursued is to build a scoring function $s(z)$, based on the training examples $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$, with a ROC curve as close as possible to ROC^*. Refer to section 2.2 for additional properties.

Therefore, as foreshadowed above, empirical performance measures in bipartite ranking should be unsurprisingly based on ranks. We propose to evaluate empirically the ranking performance of any scoring function candidate $s(z)$ in $\mathscr{S}$ by means of statistics based on the training examples $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$, of the type:

$$\widehat{W}^{\phi}_{n,m}(s) = \sum_{i=1}^{n} \phi\left(\frac{\mathrm{Rank}(s(\mathbf{X}_i))}{N+1}\right), \tag{5.2.8}$$

where $N = n+m$, $\phi : [0,1] \to \mathbb{R}$ is some Borelian nondecreasing function. This quantity is a two-sample linear rank statistic (see Definition 81) related to the score-generating function $\phi(u)$ and the samples $\{s(\mathbf{X}_1), \ldots, s(\mathbf{X}_n)\}$ and $\{s(\mathbf{Y}_1), \ldots, s(\mathbf{Y}_m)\}$. This statistic is invariant by increasing transform of the scoring function s, just like the (empirical) ROC curve and, as recalled in the previous section, it is a natural and common choice to quantify differences in distribution between the univariate samples $\{s(\mathbf{X}_1), \ldots, s(\mathbf{X}_n)\}$ and $\{s(\mathbf{Y}_1), \ldots, s(\mathbf{Y}_m)\}$, to evaluate to which extent the distribution of the first sample is stochastically larger than that of the second sample in particular. It consequently appears as legitimate to learn a scoring function s by maximizing the criterion (5.2.8). Whereas rigorous arguments are developed in Section 5.4, we highlight here that, for specific choices of the score-generating function ϕ, many relevant criteria considered in the ranking literature can be accurately approximated by statistics of this form:

- $\phi(u) = u$. The obtained statistic is the famous Wilcoxon-Mann-Whitney ranksum statistic, incidentally related to the empirical AUC, see Eq. (5.2.5).
- $\phi(u) = u\mathbb{I}\{u \geq u_0\}$, with $u_0 \in (0,1)$. It corresponds to the local AUC criterion, introduced in Clémençon and Vayatis (2007), considering the 'best' instances defined as the ones having the higher ranks.
- $\phi(u) = u^q$, with $q > 0$. Introduced in Rudin (2006) as the q-norm push, it is related to an alternative definition of the rank. This originally considers the rank of the positive instances among the negative ones, instead of the pooled sample. The study of such criterion is much simpler thanks to the independence property between the two samples, but has an increasing variance.
- $\phi(u) = \phi_N(u) = c\left((N+1)u\right)\mathbb{I}\{u \geq k/(N+1)\}$. This function is related to the DCG criterion, introduced in Cossock and Zhang (2006), and considered as a 'gold standard' quality measure in information retrieval, when *grades* are binary. The weights $c(i)$ define the *discount factors* and measure the importance of the ith rank. The integer k defines the number of top-ranked (best) instances to consider, corresponding to the higher ranks. Also, in the present context, the sequence $\{c(i)\}_{i \leq N}$ should be chosen increasing.

Depending on the choice of the score-generating function ϕ, some specific patterns of the preorder induced by a scoring function $s(z)$ can be either enhanced by the criterion (5.2.8) or else completely disappear: for instance, the value of (5.2.8) is essentially determined by the possible presence of positive instances among top-ranked observations, when considering a score generating function ϕ that rapidly vanishes near 0 and takes much higher values near 1.

Investigating the performance of maximizers of the criterion (5.2.8) from a nonasymptotic perspective is however far from straightforward, due to the complexity of the latter (*i.e.* a sum of strongly dependent random variables). It requires in particular to prove concentration inequalities for collections of two-sample linear rank statistics, indexed by classes of scoring functions of controlled complexity (*i.e.* of VC-type), referred to as two-sample rank processes throughout the chapter. It is the purpose of the next section to establish such results.

5.3 Concentration inequalities for two-sample rank processes

This section is devoted to prove concentration bounds for collections of two-sample linear rank statistics (5.2.8), indexed by classes $\mathscr{S}_0 \subset \mathscr{S}$ of scoring functions. In order to study the fluctuations of (5.2.8) as the full sample size N increases, it is of course required to control the fraction of 'positive'/'negative' observations in the pooled dataset. Let $p \in (0,1)$ be the 'theoretical' fraction of

positive instances. For $N \geq 1/p$, we suppose that $n = \lfloor pN \rfloor$ and $m = \lceil (1-p)N \rceil = N - n$. Define the mixture probability distribution $F = pG + (1-p)H$. For any $s \in \mathscr{S}$, the distribution of $s(\mathbf{X})$ (*i.e.* the image of G by s) is denoted by G_s, that of $s(\mathbf{Y})$ (*i.e.* the image of H by s) by H_s. We also denote by F_s the image of distribution F by s. For simplicity, the same notations are used to mean the related cumulative distribution functions. We also introduce their statistical versions $\widehat{G}_{s,n}(t) = (1/n)\sum_{i=1}^{n} \mathbb{I}\{s(\mathbf{X}_i) \leq t\}$ and $\widehat{H}_{s,m}(t) = (1/m)\sum_{j=1}^{m} \mathbb{I}\{s(\mathbf{Y}_j) \leq t\}$ and define:

$$\widehat{F}_{s,N}(t) = (n/N)\widehat{G}_{s,n}(t) + (m/N)\widehat{H}_{s,m}(t) \; . \tag{5.3.1}$$

Since $n/N \to p$ as N tends to infinity, the quantity above is a natural estimator of the *c.d.f.* F_s. Equipped with these notations, we can write:

$$\frac{1}{n}\widehat{W}^{\phi}_{n,m}(s) = \frac{1}{n}\sum_{i=1}^{n} \phi\left(\frac{N}{N+1}\widehat{F}_{s,N}(s(\mathbf{X}_i))\right). \tag{5.3.2}$$

Hence, the statistic (5.3.2) can be naturally seen as an empirical version of the quantity defined below, around which it fluctuates.

Definition 51. *For a given score-generating function ϕ, the functional*

$$W_{\phi}(s) = \mathbb{E}[(\phi \circ F_s)(s(\mathbf{X}))] \; , \tag{5.3.3}$$

is referred to as the "W_{ϕ}-ranking performance measure".

Indeed, replacing $\widehat{F}_{s,N}(s(\mathbf{X}_i))$ in (5.3.2) by $F_s(s(\mathbf{X}_i))$ and taking next the expectation permits to recover (A.1.2). Observe in addition that, for $\phi(u) = u$, the quantity (A.1.2) is equal to AUC(s) Def. 13 a soon as the distribution F_s is continuous. The next lemma reveals that the criterion (A.1.2) can be viewed as a scalar summary of the ROC curve.

Lemma 52. *Let ϕ be a score-generating function. We have, for all s in $\mathscr{S}$,*

$$W_{\phi}(s) = \frac{1}{p}\int_0^1 \phi(u)du - \frac{1-p}{p}\int_0^1 \phi\left(p(1-\mathrm{ROC(s,\alpha)}) + (1-\mathrm{p})(1-\alpha)\right) d\alpha \; . \tag{5.3.4}$$

PROOF. Using the decomposition $F_s = pG_s + (1-p)H_s$, we are led to the following expression:

$$pW_{\phi}(s) = \int_0^1 \phi(u)\, du - (1-p)\mathbb{E}[(\phi \circ F_s)(s(\mathbf{Y}))] \; .$$

Then, using a change of variable, we get:

$$\mathbb{E}[(\phi \circ F_s)(s(\mathbf{Y}))] = \int_0^1 \phi(p(1-\mathrm{ROC(s,\alpha)}) + (1-\mathrm{p})(1-\alpha))\, \mathrm{d}\alpha \; .$$

□

As revealed by Eq. (5.3.4), a score-generating function ϕ that takes much higher values near 1 than near 0 defines a criterion (A.1.2) that mainly summarizes the behavior of the ROC curve near the origin, *i.e.* the preorder on the set of instances with highest scores.

Below, we investigate the concentration properties of the process:

$$\left\{\frac{1}{n}\widehat{W}^{\phi}_{n,m}(s) - W_{\phi}(s)\right\}_{s \in \mathscr{S}_0} . \tag{5.3.5}$$

As a first go, we prove, by means of linearization techniques, that two-sample linear rank statistics can be uniformly approximated by much simpler quantities, involving *i.i.d.* averages and two-sample U-statistics. This will be key to establish probability bounds for the maximal deviation:

$$\sup_{s\in\mathscr{S}_0}\left|\frac{1}{n}\widehat{W}^{\phi}_{n,m}(s)-W_{\phi}(s)\right|, \tag{5.3.6}$$

under adequate complexity assumptions for the class $\mathscr{S}_0$ of scoring functions considered and to study next the generalization ability of maximizers of the empirical criterion (5.3.2) in terms of W_ϕ-ranking performance. Throughout the chapter, all the suprema considered, such as (5.3.6), are assumed to be measurable and we refer to Chapter 2.3 in van der Vaart and Wellner (1996) for more details on the formulation in terms of outer measure/expectation that guarantees measurability.

Uniform approximation of two-sample linear rank statistics. Whereas statistical guarantees for Empirical Risk Minimization in the context of classification or regression can be directly obtained by means of classic concentration results for empirical processes (*i.e.* averages of *i.i.d.* random variables), the study of the fluctuations of the process (5.3.5) is far from straightforward, insofar as the terms averaged in (5.3.2) are not independent. For averages of non-*i.i.d.* random variables, the underlying statistical structure can be revealed by orthogonal projections onto the space of sums of *i.i.d.* random variables in many situations. This projection argument was the key for the study of empirical AUC maximization or that of within cluster point scatter, which involved U-processes, see Clémençon et al. (2008) and Clémençon (2014). In the case of U-statistics, this orthogonal decomposition is known as the *Hoeffding decomposition* and the remainder may be expressed as a degenerate U-statistic, see Hoeffding (1948). For rank statistics, a similar though more complex decomposition can be considered. We refer to Hájek (1968) for a systematic use of the *projection method* for investigating the asymptotic properties of general statistics. From the perspective of ERM in statistical learning theory, through the *projection method*, well-known concentration results for standard empirical processes and U-processes may carry over to more complex collections of random variables such as *two-sample linear rank processes*, as revealed by the approximation result stated below. It holds true under the following technical assumptions.

***Assumption* 3.** *Let $M > 0$. For all $s \in \mathscr{S}_0$, the random variables $s(\mathbf{X})$ and $s(\mathbf{Y})$ are continuous, with density functions that are twice differentiable and have Sobolev $\mathscr{W}^{2,\infty}$-norms[1] bounded by $M < +\infty$.*

***Assumption* 4.** *The score-generating function $\phi : [0,1] \mapsto \mathbb{R}$, is nondecreasing and twice continuously differentiable.*

***Assumption* 5.** *The class of scoring functions $\mathscr{S}_0$ is a* VC *class of finite* VC *dimension $\mathscr{V} < +\infty$.*

For the definition of VC classes of functions, one may refer to *e.g.* van der Vaart and Wellner (1996), see section 2.6.2 therein, and also recalled in Chapter 3, Section 3.2. By means of the proposition below, the study of the fluctuations of the two-sample linear rank process (5.3.5) boils down to that of basic empirical processes.

Proposition 53. *Suppose that Assumptions 3-5 are fulfilled. The two-sample linear rank process (5.3.5) can be linearized/decomposed as follows. For all $s \in \mathscr{S}_0$,*

[1] Recall that the Sobolev space $\mathscr{W}^{2,\infty}$ is the space of all Borelian functions $h : \mathbb{R} \to \mathbb{R}$ such that h and its first and second order weak derivatives h' and h'' are bounded almost-everywhere. Denoting by $||.||_\infty$ the norm of the Lebesgue space L_∞ of Borelian and essentially bounded functions, $\mathscr{W}^{2,\infty}$ is a Banach space when equipped with the norm $||h||_{2,\infty} = \max\{||h||_\infty, ||h'||_\infty, ||h''||_\infty\}$.

$$\widehat{W}_{n,m}^{\phi}(s) = n\widehat{W}_{\phi}(s) + \left(\widehat{V}_n^X(s) - \mathbb{E}\left[\widehat{V}_n^X(s)\right]\right) + \left(\widehat{V}_m^Y(s) - \mathbb{E}\left[\widehat{V}_m^Y(s)\right]\right) + \mathscr{R}_{n,m}(s)\ , \tag{5.3.7}$$

where

$$\begin{aligned}
\widehat{W}_{\phi}(s) &= \frac{1}{n}\sum_{i=1}^{n}(\phi\circ F_s)(s(\mathbf{X}_i))\ , \\
\widehat{V}_n^X(s) &= \frac{n-1}{N+1}\sum_{i=1}^{n}\int_{s(\mathbf{X}_i)}^{+\infty}(\phi'\circ F_s)(u)dG_s(u)\ , \\
\widehat{V}_m^Y(s) &= \frac{n}{N+1}\sum_{j=1}^{m}\int_{s(\mathbf{Y}_j)}^{+\infty}(\phi'\circ F_s)(u)dG_s(u)\ .
\end{aligned}$$

For any $\delta\in(0,1)$, there exist constants c_1, $c_3>0$, $c_2\geq 1$, $c_4>6$, $c_5>3$, depending on ϕ and $\mathscr{V}$, such that

$$\mathbb{P}\left\{\sup_{s\in\mathscr{S}_0}|\mathscr{R}_{n,m}(s)|<t\right\}\geq 1-\delta\ , \tag{5.3.8}$$

where $t=c_1+c_2\log(c_4/\delta)$ as soon as $N\geq(c_3/p)\log(c_5/\delta)$.

The proof of this linearization result is detailed in the Appendix section 5.6.1 (refer to it for a description of the constants involved in the bound stated above). Its main argument consists in decomposing (5.3.2) by means of a Taylor expansion at order two of the score generating function $\phi(u)$ and applying next the Hájek orthogonal projection technique (recalled at length in the Introduction Lemma B.1 for completeness) to the component corresponding to the first order term. The quantity $\mathscr{R}_{n,m}(s)$ is then formed by bringing together the remainder of the Hájek projection and the component corresponding to the second order term of the Taylor expansion, while the probabilistic control of its order of magnitude is established by means of concentration results for (degenerate) one/two-sample U-processes (see the Appendix section 3.3.2 for more details). It follows from decomposition (6.6.1) combined with triangular inequality that:

$$\begin{aligned}
\sup_{s\in\mathscr{S}_0}\left|\frac{1}{n}\widehat{W}_{n,m}^{\phi}(s)-W_{\phi}(s)\right| &\leq \sup_{s\in\mathscr{S}_0}\left|\widehat{W}_{\phi}(s)-W_{\phi}(s)\right| \\
&+\sup_{s\in\mathscr{S}_0}\frac{1}{n}\left|\widehat{V}_n^X(s)-\mathbb{E}\left[\widehat{V}_n^X(s)\right]\right|+\sup_{s\in\mathscr{S}_0}\frac{1}{n}\left|\widehat{V}_m^Y(s)-\mathbb{E}\left[\widehat{V}_m^Y(s)\right]\right| \\
&+\sup_{s\in\mathscr{S}_0}\frac{1}{n}|\mathscr{R}_{n,m}(s)|\ .
\end{aligned} \tag{5.3.9}$$

Hence, nonasymptotic bounds for the maximal deviation of the process (5.3.5) can be deduced from concentration inequalities for standard empirical processes, as shall be seen below. Before this, a few comments are in order.

Remark 3. (ON THE COMPLEXITY ASSUMPTION) *We point out that alternative complexity measures could be naturally considered, such as those based on Rademacher averages, see e.g. Koltchinskii (2006). However, as different types of stochastic process (i.e. empirical process, degenerate one-sample U-process and degenerate two-sample U-process) are involved in the present nonasymptotic study, different types of Rademacher complexities (see e.g. Clémençon et al. (2008)) should be introduced to control their fluctuations as well. For the sake of simplicity, the concept of* VC*-type class of functions is used here.*

Remark 4. (SMOOTH SCORE-GENERATING FUNCTIONS) *The subsequent analysis is restricted to the case of smooth score-generating functions for simplification purposes. We nevertheless point out that, although one may always build smooth approximants of irregular score generating functions, the theoretical results established below can be directly extended to non-smooth situations, at the price of a significantly greater technical complexity.*

The theorem below provides a concentration bound for the two-sample rank process (5.3.5). The proof is based on the uniform approximation result precedingly established, refer to the Appendix section 5.6.3 for technical details.

Theorem 54. *Suppose that the assumptions of Proposition 53 are fulfilled. Then, there exist constants C_1, $C_3 > 0$, $C_2 \geq 24$, depending on ϕ, $\mathscr{V}$ and $C_4 \geq C_1$ depending on ϕ, such that:*

$$\mathbb{P}\left\{ \sup_{s \in \mathscr{S}_0} \left| \frac{1}{n} \widehat{W}_{n,m}^{\phi}(s) - W_{\phi}(s) \right| > t \right\} \leq C_2 e^{-pC_3 N t^2} , \tag{5.3.10}$$

as soon as $C_1/\sqrt{pN} \leq t \leq C_4 \min(p, 1-p)$.

The concentration inequalities stated above are extensively used in the next section to study the ranking bipartite learning problem, when formulated as W_ϕ-ranking performance maximization.

5.4 Performance of maximizers of two-sample rank statistics in bipartite ranking

This section provides a theoretical analysis of bipartite ranking methods, based on maximization of the empirical ranking performance measure (5.2.8). While the concentration inequalities established in the previous section are the key technical tools to derive nonasymptotic bounds for the deficit of W_ϕ-ranking performance measure of empirical maximizers, we start by showing that the criterion (A.1.2) is relevant to measure ranking performance, whatever the score generating function ϕ is chosen, beyond the examples listed in Subsection 5.2.2.

Optimal elements. The next result states that optimal scoring functions do maximize the W_ϕ-ranking performance and form a collection that coincides with the set $\mathscr{S}_\phi^*$ of maximizers of (A.1.2), provided that the score-generating function ϕ is strictly increasing on $(0,1)$.

Proposition 55. *Let ϕ be a score-generating function. The assertions below hold true.*

(i) For all $(s, s^*) \in \mathscr{S} \times \mathscr{S}^*$, we have $W_\phi(s) \leq W_\phi(s^*) = W_\phi^*$, where $W_\phi^* \overset{def}{=} W_\phi(\Psi)$.

(ii) Assuming in addition that the score-generating function ϕ is strictly increasing on $(0,1)$, we have: $\mathscr{S}_\phi^* = \mathscr{S}^*$.

The proof immediately results from (5.3.4) combined with the fact that the ROC curve of increasing transforms of the likelihood ratio $\Psi(z)$ dominates everywhere any other ROC curve, as recalled in Sections 5.2.2 and 2.2: $\forall (s,\, s^*) \in \mathscr{S} \times \mathscr{S}^*$, $\forall \alpha \in (0,\, 1)$, $\mathrm{ROC}(s, \alpha) \leq \mathrm{ROC}(s^*, \alpha) = \mathrm{ROC}^*(\alpha)$. Details are left to the reader.

Remark 5. (ON PLUG-IN RANKING RULES) *Theoretically, a possible approach to bipartite ranking is the* plug-in *method (Devroye et al. (1996)), which consists of using an estimate $\hat{\Psi}$ of the likelihood function as a scoring function. As shown by the subsequent bound, when ϕ is differentiable with a*

bounded derivative, when $\widehat{\Psi}$ is close to Ψ in the L_1-sense, it leads to a nearly optimal ordering in terms of W-ranking criterion:

$$W_\phi^* - W_\phi\left(\widehat{\Psi}\right) \leq (1-p)||\phi'||_\infty \mathbb{E}[|\widehat{\Psi}(\mathbf{X}) - \Psi(\mathbf{X})|] \,.$$

However, the bound above may be loose and the plug-in approach faces computational difficulties when dealing with high-dimensional data, see Györfi et al. (2002), which provide the motivation for exploring algorithms based on W_ϕ-ranking performance maximization.

Remark 6. (ALTERNATIVE PROBABILISTIC FRAMEWORK) *We point out that the present analysis can be extended to the alternative setup, where, rather than assuming that two samples of sizes n and m, 'positive' and 'negative', are available for the learning tasks considered in this chapter, the i.i.d. observations Z are supposed to come with a random label Y either equal to $+1$ or else to -1, indicating whether Z is distributed according to G or H. If p denotes the probability that the label Y is equal to 1, the number n of positive observations among a training sample of size N is then random, distributed as a binomial of size N with parameter p.*

Consider any maximizer of the empirical W_ϕ-ranking performance measure over a class $\mathscr{S}_0 \subset \mathscr{S}$ of scoring rules:

$$\widehat{s} \in \operatorname*{arg\,max}_{s \in \mathscr{S}_0} \widehat{W}_{n,m}^\phi(s) \,. \tag{5.4.1}$$

Since we obviously have:

$$W_\phi^* - W_\phi(\widehat{s}) \leq 2 \sup_{s \in \mathscr{S}_0} \left| \frac{1}{n} \widehat{W}_{n,m}^\phi(s) - W_\phi(s) \right| + \left(W_\phi^* - \sup_{s \in \mathscr{S}_0} W_\phi(s) \right), \tag{5.4.2}$$

the control of deficit of W-ranking performance of empirical maximizers of (5.3.2) can be deduced from the concentration properties of the process (5.3.5).

5.4.1 Generalization error bounds and model selection

The corollary below describes the generalization capacity of scoring rules based on empirical maximization of W_ϕ-ranking performance criteria. It straightforwardly results from Theorem 54 combined with the bound (5.4.2).

Corollary 56. *Let $\hat{s}$ be an empirical W_ϕ-ranking performance maximizer over the class $\mathscr{S}_0$, i.e. $\hat{s} \in \arg\max_{s \in \mathscr{S}_0} \widehat{W}_{n,m}^\phi(s)$. Under the assumptions of Proposition 53, for any $\delta \in (0,1)$, we have with probability at least $1-\delta$:*

$$W_\phi^* - W_\phi(\hat{s}) \leq 2C_3 \sqrt{\frac{\log(C_2/\delta)}{pN}} + \left(W_\phi^* - \sup_{s \in \mathscr{S}_0} W_\phi(s) \right), \tag{5.4.3}$$

as soon as $N \geq c/(p \min(p, 1-p)^2) \log(C_2/\delta)$ and $\delta \leq C_2 e^{-(C_1/C_3)^2}$ with $c > 0$ depending on ϕ, $\mathscr{V}$, and where the constants C_i, $i \leq 3$, being the same as those involved in Theorem 54.

The result above establishes that maximizers of the empirical criterion (5.2.8) achieve a classic learning rate bound of order $O_{\mathbb{P}}(1/\sqrt{N})$ when based on a training data set of size N, just like in standard classification, see *e.g.* Devroye et al. (1996). Refer to the Appendix section 5.6.4 for the proof of an additional result, that provides a bound in expectation for the deficit of W_ϕ-ranking performance measure, similar to that established in the subsequent analysis, devoted to the model selection issue.

Model selection by complexity penalization. We have investigated the issue of approximately recovering the best scoring rule in a given class $\mathscr{S}_0$ in the sense of the W_ϕ-ranking performance measure (A.1.2), which is satisfactory only when the minimum achieved over $\mathscr{S}_0$ is close to W_ϕ^* of course. We now address the problem of model selection, that is the problem of selecting a good scoring function from one of a collection of VC classes $\mathscr{S}_k$, $k \geq 1$. A model selection method is a data-based procedure that aims at achieving a trade-off regarding two contradictory objectives, *i.e.* at finding a class $\mathscr{S}_k$ rich enough to include a reasonable approximant of an element of $\mathscr{S}^*$, while being not too complex so that the performance of the empirical minimizer over it $\hat{s}_k = \arg\max_{s \in \mathscr{S}_k} \widehat{W}_{n,m}^{\phi}(s)$ can be statistically guaranteed. We suppose that all class candidates $\mathscr{S}_k$, $k \geq 1$, fulfill the assumptions of Proposition 53 and denote by $\mathscr{V}_k$ the VC dimension of the class $\mathscr{S}_k$. Various model selection techniques, based on (re-)sampling or data-splitting procedures, could be naturally considered for this purpose. Here, in order to avoid overfitting, we focus on a complexity regularization approach, of which study can be directly derived from the rate bound analysis previously carried out, that consists in substracting to the empirical ranking performance measure the penalty term (increasing with $\mathscr{V}_k$) given by:

$$\mathrm{pen}(N,k) = B_1 \sqrt{\frac{\mathscr{V}_k}{pN}} + \sqrt{\frac{2C \log k}{p^2 N}} \,, \tag{5.4.4}$$

for $pN \geq B_2 \mathscr{V}_k$ where the constants B_1 and B_2 are those involved in Proposition 63 and $C = 6(\|\phi\|_\infty^2 + 9\|\phi'\|_\infty^2 + 9\|\phi''\|_\infty^2)$. The scoring function selected maximizes the penalized empirical ranking performance measure, it is $\hat{s}_{\hat{k}}(z)$ where:

$$\hat{k} = \arg\max_{k \geq 1} \left\{ \frac{1}{n} \widehat{W}_{n,m}^{\phi}(s) - \mathrm{pen}(N,k) \right\} \,. \tag{5.4.5}$$

The result below shows that the scoring rule $\hat{s}_{\hat{k}}$ nearly achieves the expected deficit of W_ϕ-ranking performance that would have been attained with the help of an oracle, revealing the model minimizing $W_\phi^* - \mathbb{E}[W_\phi(\hat{s}_k)]$.

Proposition 57. *Suppose that the assumptions of Proposition 53 are fulfilled for any class $\mathscr{S}_k$ with $k \geq 1$ and that $\sup_{k \geq 1} \mathscr{V}_k < +\infty$. Then, we have:*

$$W_\phi^* - \mathbb{E}\left[W_\phi(\hat{s}_{\hat{k}})\right] \leq \min_{k \geq 1} \left\{ 2pen(N,k) + \left(W_\phi^* - \sup_{s \in \mathscr{S}_k} W_\phi(s) \right) \right\} + 2\sqrt{\frac{C}{p^2 N}} \,, \tag{5.4.6}$$

as soon as $pN \geq B_2 \sup_{k \geq 1} \mathscr{V}_k$, where the constant $B_2 > 0$ is the same as that involved in Proposition 63 and $C = 6(\|\phi\|_\infty^2 + 9\|\phi'\|_\infty^2 + 9\|\phi''\|_\infty^2)$.

Refer to the Appendix section 5.6.5 for the technical proof.

5.4.2 Kernel regularization for ranking performance maximization

Many successful algorithmic approaches to statistical learning (*e.g.* boosting, support vector machines, neural networks) consist in smoothing the empirical risk/performance functional to be optimized, so as to use computationally feasible techniques based on gradient descent/ascent methods. Concerning the empirical criterion (5.2.8), although one may choose a regular score generating function ϕ (*cf* Remark 4), smoothness issues arise when replacing F_s in (A.1.2) by the raw empirical *c.d.f.* (A.1.1). A classic remedy involves using a kernel-smoothed version of the empirical *c.d.f.* instead. Let $K : \mathbb{R} \to \mathbb{R}$ be a second-order Parzen-Rosenblatt kernel *i.e.* a Borelian symmetric function,

integrable *w.r.t.* the Lebesgue measure such that $\int K(t)dt = 1$ and $\int t^2 K(t)dt < +\infty$. Precisely, for any $h > 0$ and all $t \in \mathbb{R}$, define the smoothed approximation of the *c.d.f.* $F_s(t)$:

$$\widetilde{F}_{s,h}(t) = \int_{\mathbb{R}} \kappa\left(\frac{t-u}{h}\right) F_s(du) \ , \tag{5.4.7}$$

where $\kappa(t) = \int_{-\infty}^{t} K(u)du$ and $h > 0$ is the bandwidth that determines the degree of smoothing, see *e.g.* Nadaraya (1964). The uniform integrated error $\sup_{s\in\mathscr{S}_0} \int |\widetilde{F}_{s,h}(t) - F_s(t)|dt$ is shown to be of order $O(h^2)$ under the assumptions recalled below, see Jones (1990).

***Assumption* 6.** *Let $R > 0$. For all s in $\mathscr{S}_0$, the cumulative distribution function F_s is differentiable with derivative f_s such that $\int (f_s'(t))^2 dt \leq R$.*

***Assumption* 7.** *The kernel function K is of the form $K_1 \circ K_2$, where K_1 is a function of bounded variation and K_2 is a polynomial.*

Notice that Assumption 6 is fulfilled as soon as Assumption 3 is satisfied with $R \geq M$. The statistical counterpart of (5.4.7) is then:

$$\widehat{F}_{s,N,h}(t) = \frac{1}{N}\sum_{i=1}^{n} \kappa\left(\frac{t - s(\mathbf{X}_i)}{h}\right) + \frac{1}{N}\sum_{j=1}^{m} \kappa\left(\frac{t - s(\mathbf{Y}_j)}{h}\right). \tag{5.4.8}$$

A smooth version of the theoretical criterion (A.1.2) is given by:

$$\widetilde{W}_{\phi,h}(s) = \mathbb{E}[(\phi \circ \widetilde{F}_{s,h})(s(\mathbf{X}))] \ , \tag{5.4.9}$$

for all $s \in \mathscr{S}$ and an empirical version of the latter is $\widehat{W}^{\phi}_{n,m,h}(s)/n$, where:

$$\widehat{W}^{\phi}_{n,m,h}(s) = \sum_{i=1}^{n} (\phi \circ \widehat{F}_{s,N,h})(s(\mathbf{X}_i)) \ . \tag{5.4.10}$$

For any maximizer $\tilde{s}$ of (5.4.10) over the class $\mathscr{S}_0$ of scoring function candidates, we almost-surely have:

$$W^*_{\phi} - W_{\phi}(\tilde{s}) \leq 2 \sup_{s\in\mathscr{S}_0} \left| \frac{1}{n}\widehat{W}^{\phi}_{n,m,h}(s) - \widetilde{W}_{\phi,h}(s) \right| + \sup_{s\in\mathscr{S}_0} \left| \widetilde{W}_{\phi,h}(s) - W_{\phi}(s) \right| + \left\{ W^*_{\phi} - \sup_{s\in\mathscr{S}_0} W_{\phi}(s) \right\} . \tag{5.4.11}$$

This decomposition is similar to that obtained in (5.4.2) for maximizers of the criterion (5.2.8), apart from the additional bias term. Since the latter can be shown to be of order $O(h^2)$ under appropriate regularity conditions and the first term on the right hand side of the equation above can be controlled like in Theorem 54, one may bound the deficit of W_{ϕ}-ranking performance measure of $\tilde{s}$ as follows.

Proposition 58. *Suppose that the assumptions of Proposition 53 are fulfilled, as well as Assumptions 6 and 7. Let $\tilde{s}$ be any maximizer of the smoothed criterion (5.4.10) over the class $\mathscr{S}_0$. Then, for any $\delta \in (0,1)$, there exist constants C_1, $C_3 > 0$, $C_2 \geq 24$ depending on ϕ, K, R, $\mathscr{V}$ and $C_4 > 0$ is a constant depending on ϕ, K and R, such that we have with probability at least $1 - \delta$:*

$$W^*_{\phi} - W_{\phi}(\tilde{s}) \leq 2C_3\sqrt{\frac{\log(C_2/\delta)}{pN}} + C_4 h^2 + \left\{ W^*_{\phi} - \sup_{s\in\mathscr{S}_0} W_{\phi}(s) \right\}, \tag{5.4.12}$$

as soon as $N \geq (C_1/p^2 c)\log(C_2/\delta)$ and $\delta \leq C_2 e^{-(C_1/C_3)^2}$ with $c > 0$ depending on ϕ, K, R, $\mathscr{V}$.

The proof is detailed in the Appendix section 5.6.6.

5.5 Conclusion

This chapter argues that two-sample linear rank statistics provide a very flexible and natural class of empirical performance measures for bipartite ranking. We have showed that it encompasses in particular well-known criteria used in medical diagnosis and information retrieval and proved that, in expectation, these criteria are maximized by optimal scoring functions and put the emphasis on specific parts of their ROC curves, depending on the score generating function involved in the criterion considered. We have established concentration results for collections of such statistics, referred to as *two-sample rank processes* here, under general assumptions and have deduced from them statistical learning guarantees for the maximizers of such ranking criteria in the form of a generalization bound of order $O_{\mathbb{P}}(1/\sqrt{N})$, where N means the size of the pooled training sample. Algorithmic issues concerning practical maximization have also been investigated and we display numerical results supporting the theoretical analysis carried out in Chapter 7.

5.6 Proofs

The proofs of the results stated in the main corpus are detailed below.

5.6.1 Proof of Proposition 53

Let $\theta_0 \in (0,1)$. Since $\phi(u) \in \mathscr{C}^2([0,1],\mathbb{R})$ by virtue of Assumption 4, a Taylor expansion of order two yields: for all $\theta \in (0,1)$

$$\phi(\theta) = \phi(\theta_0) + (\theta - \theta_0)\phi'(\theta_0) + \int_{\theta_0}^{\theta} (\theta - u)\phi''(u)du \; . \tag{5.6.1}$$

Let $s \in \mathscr{S}_0$. For all $t \in \mathbb{R}$, we have

$$\phi\left(\frac{N\widehat{F}_{s,N}(t)}{N+1}\right) = \phi \circ F_s(t) + \left(\frac{N\widehat{F}_{s,N}(t)}{N+1} - F_s(t)\right)\phi' \circ F_s(t) + \int_{F_s(t)}^{N\widehat{F}_{s,N}(t)/(N+1)} \left(\frac{N\widehat{F}_{s,N}(t)}{N+1} - u\right)\phi''(u)du \; , \tag{5.6.2}$$

with probability one. Let $i \leq n$, for $t = s(\mathbf{X}_i)$, (5.6.2) writes:

$$\phi\left(\frac{N\widehat{F}_{s,N}(s(\mathbf{X}_i))}{N+1}\right) = \phi \circ F_s(s(\mathbf{X}_i)) + \left(\frac{N\widehat{F}_{s,N}(s(\mathbf{X}_i))}{N+1} - F_s(s(\mathbf{X}_i))\right)\phi' \circ F_s(s(\mathbf{X}_i)) + t_i(s) \quad a.s. \; , \tag{5.6.3}$$

where

$$|t_i(s)| \leq (\|\phi''\|_\infty/2)\left(N/(N+1)\widehat{F}_{s,N}(s(\mathbf{X}_i)) - F_s(s(\mathbf{X}_i))\right)^2 .$$

Hence, by summing over $i \in \{1, \ldots, n\}$, one gets that the approximation of $\widehat{W}_{n,m}(s)$ stated below holds true almost-surely:

$$\widehat{W}_{n,m}(s) = n\widehat{W}_\phi(s) + B_{n,m}(s) + \widehat{T}_{n,m}(s) \; , \tag{5.6.4}$$

where

$$B_{n,m}(s) = \sum_{i=1}^{n} \left(\frac{N\widehat{F}_{s,N}(s(\mathbf{X}_i))}{N+1} - F_s(s(\mathbf{X}_i)) \right) \phi' \circ F_s(s(\mathbf{X}_i)), \tag{5.6.5}$$

$$|\widehat{T}_{n,m}(s)| = \sum_{i=1}^{n} |t_i(s)| \leq \frac{\|\phi''\|_\infty}{2} \sum_{i=1}^{n} \left(\frac{N\widehat{F}_{s,N}(s(\mathbf{X}_i))}{N+1} - F_s(s(\mathbf{X}_i)) \right)^2 . \tag{5.6.6}$$

Linearization of $B_{n,m}(\cdot)$. First, observe that

$$\begin{aligned} B_{n,m}(s) = \frac{1}{N+1} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \mathbb{I}\{s(\mathbf{X}_j) \leq s(\mathbf{X}_i)\} \phi' \circ F_s(s(\mathbf{X}_i)) \\ + \frac{1}{N+1} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{I}\{s(\mathbf{Y}_j) \leq s(\mathbf{X}_i)\} \phi' \circ F_s(s(\mathbf{X}_i)) \\ + \sum_{i=1}^{n} \left(\frac{1}{N+1} - F_s(s(\mathbf{X}_i)) \right) \phi' \circ F_s(s(\mathbf{X}_i)) \, . \end{aligned} \tag{5.6.7}$$

Notice that the first two terms are U-processes indexed by $\mathscr{S}_0$, *cf* Chap. Section 3.3, while the last term is an empirical process. Indeed, one may write

$$B_{n,m}(s) = \frac{n(n-1)}{N+1} U_n(k_s) + \frac{nm}{N+1} U_{n,m}(\ell_s) + \widehat{K}_{n,m}(s) \, , \tag{5.6.8}$$

where

$$U_n(k_s) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \mathbb{I}\{s(\mathbf{X}_j) \leq s(\mathbf{X}_i)\} \phi' \circ F_s(s(\mathbf{X}_i)) \tag{5.6.9}$$

is a (nondegenerate) 1-sample U-process of degree 2 based on the random sample $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ with nonsymmetric kernel $k_s(x,x') = \mathbb{I}\{s(x') \leq s(x)\} \phi' \circ F_s(s(x))$ on $\mathscr{X} \times \mathscr{X}$,

$$U_{n,m}(\ell_s) = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{I}\{s(\mathbf{Y}_j) \leq s(\mathbf{X}_i)\} \phi' \circ F_s(s(\mathbf{X}_i)) \tag{5.6.10}$$

is a (nondegenerate) two-sample U-process of degree $(1,1)$ based on the samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$ with kernel $\ell_s(x,y) = \mathbb{I}\{s(y) \leq s(x)\} \phi' \circ F_s(s(x))$ on $\mathscr{X} \times \mathscr{Y}$, and

$$\widehat{K}_{n,m}(s) = \sum_{i=1}^{n} \left(\frac{1}{N+1} - F_s(s(\mathbf{X}_i)) \right) \phi' \circ F_s(s(\mathbf{X}_i))$$

is an empirical process based on the $\mathbf{X}_i$'s. In order to write $B_{n,m}$ as an empirical process plus a (negligible) remainder term, the Hoeffding decomposition is applied to the U-processes above, *cf* Appendix 3.3.1:

$$U_n(k_s) = \mathbb{E}[U_n(k_s)] + \widehat{U}_n(k_s) + \mathscr{R}_n(k_s) \, , \tag{5.6.11}$$

$$U_{n,m}(\ell_s) = \mathbb{E}[U_{n,m}(\ell_s)] + \widehat{U}_{n,m}(\ell_s) + \mathscr{R}_{n,m}(\ell_s) \, , \tag{5.6.12}$$

where

$$\widehat{U}_n(k_s) = \frac{1}{n} \sum_{i=1}^{n} k_{s,1,1}(\mathbf{X}_i) + \frac{1}{n} \sum_{i=1}^{n} k_{s,1,2}(\mathbf{X}_i) \, , \tag{5.6.13}$$

with $k_{s,1,1}(x) = \mathbb{E}[k_s(x,\mathbf{X})] - \mathbb{E}[U_n(k_s)]$ and $k_{s,1,2}(x) = \mathbb{E}[k_s(\mathbf{X},x)] - \mathbb{E}[U_n(k_s)]$, and

$$\widehat{U}_{n,m}(\ell_s) = \frac{1}{m}\sum_{j=1}^{m} \ell_{s,1,1}(\mathbf{Y}_j) + \frac{1}{n}\sum_{i=1}^{n} \ell_{s,1,2}(\mathbf{X}_i) \ , \tag{5.6.14}$$

with $\ell_{s,1,1}(y) = \mathbb{E}[\ell_s(\mathbf{X},y)] - \mathbb{E}[U_{n,m}(\ell_s)]$ and $\ell_{s,1,2}(x) = \mathbb{E}[\ell_s(x,\mathbf{Y})] - \mathbb{E}[U_{n,m}(\ell_s)]$.
Consequently, the Hájek projection of the process $B_{n,m}(s)$ is given by

$$\widehat{B}_{n,m}(s) - \mathbb{E}[\widehat{B}_{n,m}(s)] = \frac{n(n-1)}{N+1}\widehat{U}_n(k_s) + \frac{nm}{N+1}\widehat{U}_{n,m}(\ell_s) + \widehat{K}_{n,m}(s) - \mathbb{E}[\widehat{K}_{n,m}(s)] \ . \tag{5.6.15}$$

The following result provides an approximation of (5.6.15) and is proved in Appendix 5.6.2.

Lemma 59. *Under Assumptions 3-5, the Hájek projection of the stochastic process $B_{n,m}(\cdot)$, denoted by $\widehat{B}_{n,m}(\cdot)$ and indexed by $\mathscr{S}_0$, onto the subspace generated by the random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ can be approximated as follows: for all $s \in \mathscr{S}_0$,*

$$\widehat{B}_{n,m}(s) - \mathbb{E}\left[\widehat{B}_{n,m}(s)\right] = \widehat{V}_n^X(s) + \widehat{V}_m^Y(s) + \widehat{R}_{n,m}(s) \ , \tag{5.6.16}$$

where

$$\widehat{V}_n^X(s) = \frac{n-1}{N+1}\sum_{i=1}^{n} k_{s,1,1}(\mathbf{X}_i), \ \widehat{V}_m^Y(s) = \frac{n}{N+1}\sum_{j=1}^{m} \ell_{s,1,1}(\mathbf{Y}_j) \ .$$

Let $\delta > 0$, there exist constants $A_1, A_3 > 0$, $A_2 \geq 1$ depending on ϕ and $\mathscr{V}$ and for all $A_4 \geq A_1$, such that

$$\mathbb{P}\left\{\sup_{s\in\mathscr{S}_0}\left|\widehat{R}_{n,m}(s)\right| > t\right\} \leq A_2 \exp\left\{-\frac{A_3 N t^2}{p\sigma^2}\right\} \ , \tag{5.6.17}$$

as soon as $A_1\sigma\sqrt{p\log(2\|\phi'\|_\infty/\sigma)/N} \leq t \leq 2pA_4\|\phi'\|_\infty$, with $\sigma^2 = \int_{[0,1]}\phi'^2$.

The last step relies on all previous decompositions, so as to approximate $B_{n,m}(\cdot)$ by the sum of two empirical processes $\widehat{V}_n^X(\cdot)$ and $\widehat{V}_n^Y(\cdot)$, with a uniform control of the error. All residual terms, $\widehat{R}_{n,m}(s)$ (Lemma 59) plus the remainders of the U-processes, are the components of the process $\mathscr{R}_{n,m}^B(s)$, see the following Lemma 60.

Lemma 60. *Suppose that Assumptions 3-5 are fulfilled. The stochastic process $B_{n,m}(.)$ can be approximated as follows: for all $s \in \mathscr{S}_0$,*

$$B_{n,m}(s) - \mathbb{E}\left[B_{n,m}(s)\right] = \widehat{V}_n^X(s) + \widehat{V}_m^Y(s) + \mathscr{R}_{n,m}^B(s) \ . \tag{5.6.18}$$

Let $\delta > 0$. There exist $D_1 > 0$ universal constant, and constants D_3, $D_4 > 0$, $D_2 \geq 1$, d_1, $d_2 > 3$ depending on ϕ and $\mathscr{V}$, such that with probability at least $1-\delta$:

$$\sup_{s\in\mathscr{S}_0}|\mathscr{R}_{n,m}^B(s)| \leq \|\phi'\|_\infty\sqrt{p(1-p)D_1\log(d_1/\delta)} + (p\|\phi'\|_\infty D_4)\log(d_2/\delta) \ , \tag{5.6.19}$$

as soon as $N \geq (pD_3)^{-1}\log(D_2/\delta)$.

Refer to Appendix 5.6.2 for the detailed proof.

A uniform bound for $\widehat{T}_{n,m}(\cdot)$. By virtue of (5.6.5), we have:

$$\sup_{s\in\mathscr{S}_0}|\widehat{T}_{n,m}(s)|\leq n\|\phi''\|_\infty\left(\sup_{(s,t)\in\mathscr{S}_0\times\mathbb{R}}\left(\widehat{F}_{s,N}(t)-F_s(t)\right)^2+\frac{1}{(N+1)^2}\right). \tag{5.6.20}$$

Observe also that

$$\sup_{(s,t)\in\mathscr{S}_0\times\mathbb{R}}|\widehat{F}_{s,N}(t)-F_s(t)|\leq p\sup_{(s,t)\in\mathscr{S}_0\times\mathbb{R}}|\widehat{G}_{s,n}(t)-G_s(t)| + (1-p)\sup_{(s,t)\in\mathscr{S}_0\times\mathbb{R}}|\widehat{H}_{s,m}(t)-H_s(t)|+\frac{2}{N}. \tag{5.6.21}$$

A classic concentration bound for empirical processes based on the VC inequality (see *e.g.* Theorems 3.2 and 3.4 in Boucheron et al. (2005)) shows that, for any $\delta\in(0,1)$, we have with probability at least $1-\delta$:

$$\sup_{(s,t)\in\mathscr{S}_0\times\mathbb{R}}|\widehat{G}_{s,n}(t)-G_s(t)|\leq c\sqrt{\frac{\mathscr{V}}{n}}+\sqrt{\frac{2\log(1/\delta)}{n}},$$

where $c>0$ is a universal constant. In a similar fashion, we have, with probability larger than $1-\delta$,

$$\sup_{(s,t)\in\mathscr{S}_0\times\mathbb{R}}|\widehat{H}_{s,m}(t)-H_s(t)|\leq c\sqrt{\frac{\mathscr{V}}{m}}+\sqrt{\frac{2\log(1/\delta)}{m}}.$$

Combining the bounds above with the union bound, (5.6.21) and (5.6.20) we obtain that, for any $\delta\in(0,1)$, we have with probability larger than $1-\delta$:

$$\sup_{s\in\mathscr{S}_0}|\widehat{T}_{n,m}(s)|\leq n||\phi''||_\infty\left(12\left(\frac{c^2\mathscr{V}+\log(2/\delta)}{N}+\frac{1}{N^2}\right)+\frac{1}{(N+1)^2}\right) \leq B_1+B_2\log(2/\delta), \tag{5.6.22}$$

where B_1 (*resp.* B_2) is a constant that only depends on ϕ and $\mathscr{V}$ (*resp.* on ϕ).
To end the proof, it suffices to observe that the remainder process is the sum of $\mathscr{R}^B_{n,m}(s)$ and $\widehat{T}_{n,m}(s)$. Combining bounds (5.6.19) and (5.6.22), we get that, with probability at least $1-\delta$,

$$\sup_{s\in\mathscr{S}_0}|\mathscr{R}_{n,m}(s)|=\sup_{s\in\mathscr{S}_0}|\mathscr{R}^B_{n,m}(s)+\widehat{T}_{n,m}(s)|\leq B_1+\|\phi'\|_\infty\kappa_pD\log(2d/\delta)+B_2\log(4/\delta) \tag{5.6.23}$$

as soon as $N\geq(pD_3)^{-1}\log(D_2/\delta)$, with $D=\max(\sqrt{D_1},D_4)$, $d=\max(d_1,d_2)$, $\kappa_p=\max(\sqrt{p(1-p)},p)$. As $B_2>1$, $d\geq 3$, and for small δ, we obtain the upperbound $B_1+(\|\phi'\|_\infty\kappa_pD+B_2)\log(2d/\delta)$.

5.6.2 Intermediary results

The intermediary results involved in Section 5.6.1 are now established.

Permanence Properties

The lemmas below claim that the collections of kernels/functions involved in the decomposition obtained in Appendix 5.6.1 are of VC-type and uniformly bounded.

Lemma 61. *Suppose that Assumptions 4 and 5 are fulfilled. Then, the collections of kernels* $\{k_s(x,x') : s \in \mathscr{S}_0\}$ *and* $\{\ell_s(x,y) : s \in \mathscr{S}_0\}$ *and are bounded* VC*-type classes of functions with parameters fully determined by* $\mathscr{V}$ *and* ϕ.

PROOF. Recall that: $\forall (x,x') \in \mathscr{X}^2$,

$$k_s(x,x') = \mathbb{I}\{s(x') \leq s(x)\} \left(\phi' \circ F_s\right)(s(x)).$$

Hence, we have $\sup_{(x,x')\in\mathscr{X}^2} |k_s(x,x')| \leq ||\phi'||_\infty$ for all $s \in \mathscr{S}_0$. In additions, since the collections $\{(x,x') \in \mathscr{X}^2 \mapsto s(x) : s \in \mathscr{S}_0\}$ and $\{(x,x') \in \mathscr{X}^2 \mapsto s(x') : s \in \mathscr{S}_0\}$ are VC classes of functions, classic permanence properties of VC classes of functions (see *e.g.* Lemma 2.6.18) shows that $\{(x,x') \in \mathscr{X}^2 \mapsto s(x) - s(x') : s \in \mathscr{S}_0\}$ is also a VC class, as well as the class of indicator functions $\{(x,x') \in \mathscr{X}^2 \mapsto \mathbb{I}\{s(x') \leq s(x)\} : s \in \mathscr{S}_0\}$. Consequently, the argument of Lemma 49's proof permits to see easily that $\{(x,x') \in \mathscr{X}^2 \mapsto F_s(s(x)) = \mathbb{E}[\mathbb{I}\{s(X) \leq s(x)\}] : s \in \mathscr{S}_0\}$ is of VC type, just like $\{(x,x') \in \mathscr{X}^2 \mapsto (\phi' \circ F_s)(s(x)) : s \in \mathscr{S}_0\}$ using the Lipschitz property of ϕ', *cf* Assumption 4. Finally, being composed of products of a function in the bounded VC-type class $\{(x,x') \in \mathscr{X}^2 \mapsto \mathbb{I}\{s(x') \leq s(x)\} : s \in \mathscr{S}_0\}$ by a function in the bounded VC-type class $\{(x,x') \in \mathscr{X}^2 \mapsto (\phi' \circ F_s)(s(x)) : s \in \mathscr{S}_0\}$, the collection $\{k_s : s \in \mathscr{S}_0\}$ is still a bounded VC-type class of functions. A similar reasoning can be applied to show that $\{\ell_s : s \in \mathscr{S}_0\}$ is a bounded VC-type class of kernels on $\mathscr{X} \times \mathscr{Y}$. □

The following result is straightforwardly deduced from the lemma above combined with Lemma 49.

Lemma 62. *Suppose that Assumptions 4 and 5 are fulfilled. Then, the collections of functions/kernels* $\{k_{s,1,1}(x) : s \in \mathscr{S}_0\}$, $\{k_{s,1,2}(x) : s \in \mathscr{S}_0\}$, $\{k_s(x,x') - k_{s,1,1}(x) - k_{s,1,2}(x') : s \in \mathscr{S}_0\}$, $\{\ell_{s,1,1}(y) : s \in \mathscr{S}_0\}$, $\{\ell_{s,1,2}(x) : s \in \mathscr{S}_0\}$ *and* $\{\ell_s(x,y) - \ell_{s,1,1}(y) - \ell_{s,1,2}(x) : s \in \mathscr{S}_0\}$ *are bounded* VC*-type classes with parameters fully determined by* $\mathscr{V}$ *and* ϕ.

Proof of Lemma 59

For $s \in \mathscr{S}_0$, by adding the diagonal term, the empirical process can be written

$$\widehat{R}_{n,m}(s) = \left(\frac{n}{N+1} - p\right) \sum_{i=1}^{n} k_{s,1,2}(\mathbf{X}_i) + \left(\frac{m}{N+1} - (1-p)\right) \sum_{i=1}^{n} \ell_{s,1,2}(\mathbf{X}_i) \,. \tag{5.6.24}$$

We uniformly bound all three empirical processes in probability using classic concentration bounds, see *e.g.* Theorem 2.1 in Giné and Guillou (2002), as follows. Assuming Assumptions 4-5, Lemma 62 states that each class of functions $\{k_{s,1,2} : s \in \mathscr{S}_0\}$, $\{\ell_{s,1,2} : s \in \mathscr{S}_0\}$ is uniformly bounded and VC-type of parameters depending only on ϕ and on the VC dimension $\mathscr{V}$. For the class $\{x \mapsto \phi' \circ F_s(s(x)) : s \in \mathscr{S}_0\}$, the arguments are exposed in the proof of Lemma 61. The variance of the kernels can be bounded for all $s \in \mathscr{S}_0$, by $\sigma^2 = \int_{[0,1]} \phi'^2$ and $\sigma^2 \leq ||\phi'||_\infty^2$ and notice that $|n/(N+1) - p| \leq 1/N$ and $|m/(N+1) - (1-p)| \leq 1/N$. Let $t > 0$, there exist a sequence of constants $A_{1,i} > 0, A_{2,i} \geq 1, A_{3,i} > 0$ depending on ϕ and $\mathscr{V}$, $i \in \{1,2\}$, such that for all $A_{4,i} \geq A_{1,i}$, the following inequalities hold.

$$\mathbb{P}\left\{\frac{1}{N} \sup_{s\in\mathscr{S}_0} \left|\sum_{i=1}^{n} k_{s,1,2}(\mathbf{X}_i)\right| > t\right\} \leq A_{2,1} \exp\left\{-\frac{A_{3,1} N t^2}{p\sigma^2}\right\} \,, \tag{5.6.25}$$

as soon as $A_{1,1}\sigma\sqrt{p\log(2\|\phi'\|_\infty/\sigma)/N} \le t \le pA_{4,1}\|\phi'\|_\infty$,

$$\mathbb{P}\left\{\frac{1}{N}\sup_{s\in\mathscr{S}_0}\left|\sum_{i=1}^{n}\ell_{s,1,2}(\mathbf{X}_i)\right| > t\right\} \le A_{2,2}\exp\left\{-\frac{A_{3,2}Nt^2}{p\sigma^2}\right\}, \tag{5.6.26}$$

as soon as $A_{1,2}\sigma\sqrt{p\log(2\|\phi'\|_\infty/\sigma)/N} \le t \le pA_{4,2}\|\phi'\|_\infty$. The union bound with threshold $t/2$ yields

$$\mathbb{P}\left\{\sup_{s\in\mathscr{S}_0}\left|\widehat{R}_{n,m}(s)\right| > t\right\} \le A_2\exp\left\{-\frac{A_3Nt^2}{p\sigma^2}\right\}, \tag{5.6.27}$$

as soon as $A_1\sigma\sqrt{p\log(2\|\phi'\|_\infty/\sigma)/N} \le t \le 2pA_4\|\phi'\|_\infty$, with $A_1 = 2\max(A_{1,1},A_{1,2})$, $A_2 = 2\max(A_{2,1},A_{2,2})$, $A_3 = \min(A_{3,1},A_{3,2})/4$, $A_4 = \min(A_{4,1},A_{4,2})$ such that $A_4 \ge A_1$.

Proof of Lemma 60

The remainder of the decomposition (60) is obtained by combining Eq. (5.6.8), (5.6.15) and yields, for all $s \in \mathscr{S}_0$

$$\left|\mathscr{R}^B_{n,m}(s)\right| \le |\widehat{R}_{n,m}(s)| + p^2N|\mathscr{R}_n(k_s)| + p(1-p)N|\mathscr{R}_{n,m}(\ell_s)| .$$

Suppose Assumptions 4-5 are fulfilled. The first process can be uniformly bounded on $\mathscr{S}_0$ as proved in Lemma 59. For the two others, we apply the results of Lemmas 44 and 45 as follows. The process $\mathscr{R}_n(k_s)$ (*resp.* $\mathscr{R}_{n,m}(\ell_s)$) is the residual term obtained by decomposing the U-process $U_n(k_s)$ (Eq. (5.6.11), *resp.* (5.6.12))), for all $s \in \mathscr{S}_0$. By Lemma 62, its class of degenerate kernels $\{(x,x') \mapsto k_s(x,x') - k_{s,1,1}(x) - k_{s,1,2}(x') : s \in \mathscr{S}_0\}$ (*resp.* $\{(x,y) \mapsto \ell_s(x,y) - \ell_{s,1,1}(y) - \ell_{s,1,2}(x) : s \in \mathscr{S}_0\}$) is uniformly bounded and VC-type of parameters depending only on ϕ and on the VC dimension $\mathscr{V}$. Notice that the three classes of functions have variances and envelopes which can be similarly bounded by $\sigma^2 = \int_{[0,1]}\phi'^2 \le ||\phi'||^2_\infty$, up to a multiplicative constant for both residuals. Let $\delta > 0$, there exist constants $A_1, B_1 > 0, A_2, B_2 \ge 1, A_3, B_3 > 0$ depending on ϕ and $\mathscr{V}$ *s.t.* with probability at least $1-\delta$

$$\sup_{s\in\mathscr{S}_0}\left|\widehat{R}_{n,m}(s)\right| \le \|\phi'\|_\infty\sqrt{\frac{p\log(A_2/\delta)}{A_3N}}, \tag{5.6.28}$$

as soon as $N \ge (4pA_3)^{-1}\log(A_2/\delta)$. Also by Lemma 44

$$p^2N\sup_{s\in\mathscr{S}_0}|\mathscr{R}_n(k_s)| \le (p\|\phi'\|_\infty/B_3)\log(B_2/\delta), \tag{5.6.29}$$

when $N \ge (pB_3)^{-1}\log(B_2/\delta)$. And, by Lemma 45, there exist constants $C_1 > 0$, $C_2 > 1$ depending on $\mathscr{V}$, ϕ and a universal constant $C_3 > 0$ such that

$$p(1-p)N\sup_{s\in\mathscr{S}_0}|\mathscr{R}_{n,m}(\ell_s)| \le \|\phi'\|_\infty\sqrt{p(1-p)C_3\log(C_2/\delta)}, \tag{5.6.30}$$

for $\log(C_2/\delta) \ge C_1(\|\phi'\|^2_\infty C_3)^{-1}$. The union bound concludes by considering constants such that with probability at least $1-\delta$

$$\sup_{s\in\mathscr{S}_0}|\mathscr{R}^B_{n,m}(s)| \le \|\phi'\|_\infty\sqrt{p(1-p)C_3\log(3C_2/\delta)} + (p\|\phi'\|_\infty/B_3)\log(3B_2/\delta), \tag{5.6.31}$$

as soon as $N \ge (pD_3)^{-1}\log(D_2/\delta)$, where $D_2 = 3\max(A_2,B_2)$ and $D_3 = \min(4A_3,B_3)$.

5.6.3 Proof of Theorem 54

Observe, by virtue of Proposition 53 and for all $s \in \mathscr{S}_0$

$$\left|\frac{1}{n}\widehat{W}^{\phi}_{n,m}(s) - W_{\phi}(s)\right| \leq \frac{1}{n}\left|\sum_{i=1}^{n} \phi \circ F_s(s(\mathbf{X}_i)) - \mathbb{E}[\phi \circ F_s(s(\mathbf{X}))]\right| + \frac{1}{N}\left|\sum_{i=1}^{n} k_{s,1,1}(\mathbf{X}_i)\right| + \frac{1}{N}\left|\sum_{j=1}^{m} \ell_{s,1,1}(\mathbf{Y}_j)\right| + \frac{1}{n}\left|\mathscr{R}_{n,m}(s)\right| .$$

Under Assumptions 4-5, we sequentially provide uniform bounds in probability for all processes. The classes of kernels $\{x \mapsto k_{s,1,1}(x) : s \in \mathscr{S}_0\}$ and $\{y \mapsto \ell_{s,1,1}(y) : s \in \mathscr{S}_0\}$, by Lemma 62, are bounded and VC-type of parameters depending on ϕ and on the VC dimension $\mathscr{V}$ of $\mathscr{S}_0$. Their variance can be bounded, for all $s \in \mathscr{S}_0$, by $\sigma^2 = \int_{[0,1]} \phi'^2$ and $\sigma^2 \leq ||\phi'||^2_{\infty}$. As well for the collection $\{x \mapsto \phi \circ F_s(s(x)) : s \in \mathscr{S}_0\}$ where the arguments are detailed in Lemma 61 and of variance bounded by, for all $s \in \mathscr{S}_0$, by $\Sigma^2 = \int_{[0,1]} \phi^2$ and $\Sigma^2 \leq ||\phi||^2_{\infty}$. Similarly to Lemma 59, we apply Theorem 2.1 in Giné and Guillou (2002) to the empirical processes $\widehat{W}_{\phi}(s)$, $\widehat{V}^X_n(s)$ and $\widehat{V}^Y_m(s)$ as follows.
Let $t > 0$. There exist a sequence of constants $C_{1,i} > 0, C_{2,i} \geq 1, C_{3,i} > 0$ depending on ϕ and $\mathscr{V}$, such that for all $C_{4,i} \geq C_{1,i}$, $i \in \{1,2,3\}$, the following inequalities hold true.

$$\mathbb{P}\left\{\sup_{s\in\mathscr{S}_0}\left|\widehat{W}_{\phi}(s) - W_{\phi}(s)\right| > t\right\} \leq C_{2,1}\exp\left\{-\frac{C_{3,1}pNt^2}{\Sigma^2}\right\} , \tag{5.6.32}$$

as soon as $C_{1,1}\|\phi\|_{\infty}\sqrt{(1/pN)\log(2\|\phi\|_{\infty}/\Sigma)} \leq t \leq C_{4,1}\|\phi\|_{\infty}$.

$$\mathbb{P}\left\{\frac{1}{N}\sup_{s\in\mathscr{S}_0}\left|\sum_{i=1}^{n} k_{s,1,1}(\mathbf{X}_i)\right| > t\right\} \leq C_{2,2}\exp\left\{-\frac{C_{3,2}Nt^2}{p\sigma^2}\right\} , \tag{5.6.33}$$

as soon as $C_{1,2}||\phi'||_{\infty}\sqrt{(p/N)\log(2\|\phi'\|_{\infty}/\sigma)} \leq t \leq pC_{4,2}||\phi'||_{\infty}$.

$$\mathbb{P}\left\{\frac{1}{N}\sup_{s\in\mathscr{S}_0}\left|\sum_{j=1}^{m} \ell_{s,1,1}(\mathbf{Y}_j)\right| > t\right\} \leq C_{2,3}\exp\left\{-\frac{C_{3,3}Nt^2}{(1-p)\sigma^2}\right\} , \tag{5.6.34}$$

as soon as $C_{1,3}||\phi'||_{\infty}\sqrt{((1-p)/N)\log(2\|\phi'\|_{\infty}/\sigma)} \leq t \leq (1-p)C_{4,3}||\phi'||_{\infty}$. Proposition 53 provides the existence of constants $C > 6$, $D > 0$ and $c_3 > 0$, $c_5 > 3$ depending on ϕ and $\mathscr{V}$, such that

$$\mathbb{P}\left\{\frac{1}{n}\sup_{s\in\mathscr{S}_0}|\mathscr{R}_{n,m}(s)| > t\right\} \leq C\exp\left\{-\frac{pNt}{(\|\phi'\|_{\infty}\kappa_p D + B_2)}\right\} , \tag{5.6.35}$$

as soon as $N \geq (c_3/p)\log(c_5/\delta)$. The remainder process is negligible with respect to the empirical processes and we gather the four bounds to get

$$\mathbb{P}\left\{\sup_{s\in\mathscr{S}_0}\left|\frac{1}{n}\widehat{W}^{\phi}_{n,m}(s) - W_{\phi}(s)\right| > t\right\} \leq C_2 e^{-C_3 Nt^2} , \tag{5.6.36}$$

where $C_2 = 4\max(\{C_{2,i}, i \leq 3\}, C)$, $C_3 = (1/9)\min(C_{3,1}p/\Sigma^2, C_{3,2}/(p\sigma^2), C_{3,3}/((1-p)\sigma^2))$, as soon as (5.6.35) is satisfied and $C_1/\sqrt{pN} \leq t \leq C_4\min(p, 1-p)$, $C_1 > 0$ depending on ϕ, $\mathscr{V}$ and $C_4 \geq \max(C_{1,i},\ i \leq 3)$ depending on ϕ.

5.6.4 A generalization bound in expectation

For the sake of completeness, we state and prove a version in expectation of the generalization result formulated in Corollary 56.

Proposition 63. *Under the assumptions of Proposition 53, the expected risk bound is derived as follows:*

$$\mathbb{E}\left[W_\phi^* - W_\phi(\hat{s})\right] \leq B_1\sqrt{\frac{\mathscr{V}}{pN}} + W_\phi^* - \mathbb{E}\left[\sup_{s\in\mathscr{S}_0} W_\phi(s)\right], \tag{5.6.37}$$

for $pN \geq B_2\mathscr{V}$ with constants $B_1, B_2 > 0$ depending on ϕ, $\mathscr{V}$.

PROOF. Following the decomposition (5.3.9), we bound in expectation each process recalling that they are indexed by uniformly bounded VC-type classes, refer to Proof 5.6.3 for the details on theoretical guarantees concerning the permanence properties. For the empirical processes $\widehat{W}_\phi$, $\widehat{V}_n^X$ and $\widehat{V}_m^Y$, we use Theorem 2.1 in Giné and Guillou (2002), whereas for the remainder process, we require the following result that is proved subsequently.

Lemma 64. *Under the assumptions of Proposition 53, the remainder process can be uniformly bounded in expectation as follows:*

$$\mathbb{E}\left[\sup_{s\in\mathscr{S}_0} |\mathscr{R}_{n,m}(s)|\right] \leq D_1(1 + 1/p + 1/\sqrt{p(1-p)}), \tag{5.6.38}$$

for $pN \geq D_2\mathscr{V}$ with constants $D_1 > 0$ depending on ϕ, $\mathscr{V}$ and $D_2 > 0$ on ϕ.

By means of Giné and Guillou (2002), there exist universal constants $B_i > 0$, and $b_i > 0$, $i \in \{1, 2, 3\}$, depending on ϕ, $\mathscr{V}$ such that the inequalities below hold true.

$$\mathbb{E}\left[\sup_{s\in\mathscr{S}_0}\left|\widehat{W}_\phi(s) - W_\phi(s)\right|\right] \leq B_1\left(b_1\frac{\mathscr{V}\|\phi\|_\infty}{pN} + \|\phi\|_\infty\sqrt{b_1\frac{\mathscr{V}}{pN}}\right), \tag{5.6.39}$$

and

$$\mathbb{E}\left[\frac{1}{n}\sup_{s\in\mathscr{S}_0}\left|\widehat{V}_n^X(s) - \mathbb{E}\left[\widehat{V}_n^X(s)\right]\right|\right] \leq B_2\left(b_2\frac{\mathscr{V}\|\phi'\|_\infty}{pN} + \|\phi'\|_\infty\sqrt{b_2\frac{\mathscr{V}}{pN}}\right), \tag{5.6.40}$$

as well as

$$\mathbb{E}\left[\frac{1}{n}\sup_{s\in\mathscr{S}_0}\left|\widehat{V}_m^Y(s) - \mathbb{E}\left[\widehat{V}_m^Y(s)\right]\right|\right] \leq B_3\left(b_3\frac{\mathscr{V}\|\phi'\|_\infty}{pN} + \|\phi'\|_\infty\sqrt{b_3\frac{\mathscr{V}}{pN}}\right), \tag{5.6.41}$$

observing that $\int_{[0,1]}\phi^2 \leq \|\phi\|_\infty^2$ and $\int_{[0,1]}\phi'^2 \leq \|\phi'\|_\infty^2$. The remainder process being of higher order, we conclude

$$\mathbb{E}\left[\sup_{s\in\mathscr{S}_0}\left|\frac{1}{n}\widehat{W}_{n,m}^\phi(s) - W_\phi(s)\right|\right] \leq B\sqrt{b\frac{\mathscr{V}}{pN}}, \tag{5.6.42}$$

for $pN \geq \max(b, D_2)\mathscr{V}$ with constants $B > 0$ depending on ϕ and $b > 0$ depending on ϕ, $\mathscr{V}$.

□

PROOF. For all $s \in \mathscr{S}_0$

$$|\mathscr{R}_{n,m}(s)| \leq |\widehat{R}_{n,m}(s)| + N|\mathscr{R}_n(k_s)| + N|\mathscr{R}_{n,m}(\ell_s)| + |\widehat{T}_{n,m}(s)| \,. \tag{5.6.43}$$

The process appearing first in the remainder induced by the Hájek projection method (Lemma 59), is composed of sums of empirical processes, hence applying Theorem 2.1 in Giné and Guillou (2002) to each process of (6.6.7) yields

$$\mathbb{E}\left[\sup_{s\in\mathscr{S}_0}\left|\widehat{R}_{n,m}(s)\right|\right] \leq D_1\left(d\frac{\mathscr{V}\|\phi'\|_\infty}{N} + \|\phi'\|_\infty\sqrt{d\frac{p\mathscr{V}}{N}}\right) , \tag{5.6.44}$$

with constants $D_1 > 0$ depending on ϕ and $d > 0$ on ϕ, $\mathscr{V}$.The stochastic processes $\mathscr{R}_n(k_s)$ and $\mathscr{R}_{n,m}(\ell_s)$ being both degenerate U-processes, respectively one-sample of degree 2 and two-sample of degree $(1,1)$, we apply results in Nolan and Pollard (1987) (see Theorem 6 therein) and Neumeyer (2004) (see Lemma 2.4 therein) so as to get

$$\mathbb{E}\left[\sup_{s\in\mathscr{S}_0}|\mathscr{R}_n(k_s)|\right] \leq \frac{D_2\mathscr{V}}{pN} , \tag{5.6.45}$$

and

$$\mathbb{E}\left[\sup_{s\in\mathscr{S}_0}|\mathscr{R}_{n,m}(\ell_s)|\right] \leq \frac{D_3\mathscr{V}}{\sqrt{p(1-p)}N} , \tag{5.6.46}$$

D_2, $D_3 > 0$ constants of $\phi, \mathscr{V}$. For $\widehat{T}_{n,m}(s)$, the concentration inequality proved in Eq. (5.6.22) holds true for all $\delta \in (0,1)$. Hence, we have

$$\mathbb{E}\left[\sup_{s\in\mathscr{S}_0}\left|\widehat{T}_{n,m}(s)\right|\right] \;\leq\; u \;+\; \int_u^\infty \mathbb{P}\left\{\sup_{s\in\mathscr{S}_0}\left|\widehat{T}_{n,m}(s)\right| \geq x\right\}dx \;=\; u \;+\; 2B_2e^{-(u-B_1)/B_2} \quad . \tag{5.6.47}$$

Minimizing the bound above *w.r.t.* $u > 0$, we obtain the point $B_1 + B_2\log(2)$ and the upperbound then writes $B_1 + B_2(1+\log(2))$, where B_1 (*resp.* B_2) is a constant that only depends on ϕ and $\mathscr{V}$ (*resp.* on ϕ). Combining all bounds together permits to conclude: for $N \geq \mathscr{V}\log(d)$, we have

$$\begin{aligned}\mathbb{E}\left[\sup_{s\in\mathscr{S}_0}|\mathscr{R}_{n,m}(s)|\right] &\leq D_1\|\phi'\|_\infty + \frac{D_2\mathscr{V}}{p} + \frac{D_3\mathscr{V}}{\sqrt{p(1-p)}} + B_1 + B_2(1+\log(2)) \\ &\leq D(1+1/p+1/\sqrt{p(1-p)}) \,, \end{aligned} \tag{5.6.48}$$

where $D > 0$ constant depending on ϕ, $\mathscr{V}$.□

5.6.5 Proof of Proposition 57

We first prove the following lemma.

Lemma 65. *Let $\mathscr{S}_0 \subset \mathscr{S}$ and suppose that Assumptions 3-5 are fulfilled. For all $t > 0$, we have:*

$$\begin{aligned}\mathbb{P}\left\{\sup_{s\in\mathscr{S}_0}\left|W_\phi(s) - \widehat{W}^\phi_{n,m}(s)/n\right| \geq \mathbb{E}\left[\sup_{s\in\mathscr{S}_0}\left|W_\phi(s) - \widehat{W}^\phi_{n,m}(s)/n\right|\right] + t\right\} \\ \leq \exp\left\{-\frac{p^2Nt^2}{6(\|\phi\|^2_\infty + 9\|\phi'\|^2_\infty + 9||\phi''||^2_\infty)}\right\} \,. \end{aligned} \tag{5.6.49}$$

PROOF. Recall the decomposition of $\widehat{W}^{\phi}_{n,m}(s)$, for all $s \in \mathscr{S}_0$, proved in Proposition 53

$$\widehat{W}_{n,m}(s) = n\widehat{W}_{\phi}(s) + B_{n,m}(s) + \widehat{T}_{n,m}(s) \ . \tag{5.6.50}$$

Considering that $\sup_{s\in\mathscr{S}_0} \left| W_{\phi}(s) - \widehat{W}^{\phi}_{n,m}(s)/n \right|$ is a function of the N independent random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n, \mathbf{Y}_1, \ldots, \mathbf{Y}_m$, observe that changing the value of any of the $\mathbf{X}_i$'s while keeping all the others fixed changes the value of the supremum by at most

$$2||\phi||_{\infty} + 2||\phi'||_{\infty}\left(1 + \frac{m + 2(n-1)}{N+1}\right) + 2||\phi''||_{\infty}\frac{1+2m}{N^2} \ ,$$

taking into account the jumps of each of the three terms involved in (5.6.50), see Eq. (5.6.7) and (5.6.20). In a similar way, changing the value of any of the $\mathbf{Y}_j$'s changes the value of the supremum by at most

$$2||\phi'||_{\infty}\frac{n}{N+1} + 2||\phi''||_{\infty}\frac{1+2n}{N^2} \ .$$

When taking the squares, both can be upperbounded by $12(\|\phi\|^2_{\infty} + 9\|\phi'\|^2_{\infty} + 9||\phi''||^2_{\infty})$. The desired bound stated then straightforwardly results from the application of the bounded difference inequality, see McDiarmid (1989). □

Let $\varepsilon > 0$, using Proposition 63 and Lemma 65, we have, for any $k \geq 1$,

$$\begin{aligned}
&\mathbb{P}\left\{ \widehat{W}^{\phi}_{n,m}(\hat{s}_k) - B_1\sqrt{\frac{\mathscr{V}_k}{pN}} - W_{\phi}(\hat{s}_k) > \varepsilon \right\} \\
&\qquad \leq \mathbb{P}\left\{ \sup_{s\in\mathscr{S}_k} \left| W_{\phi}(s) - \widehat{W}^{\phi}_{n,m}(s)/n \right| > \mathbb{E}\left[\sup_{s\in\mathscr{S}_k} \left| W_{\phi}(s) - \widehat{W}^{\phi}_{n,m}(s)/n \right| \right] + \varepsilon \right\} \\
&\qquad\qquad \leq \exp\left\{ -\frac{p^2N\varepsilon^2}{C} \right\} \ , \qquad (5.6.51)
\end{aligned}$$

as soon as $pN \geq B_2\mathscr{V}_k$ and where $C = 6(\|\phi\|^2_{\infty} + 9\|\phi'\|^2_{\infty} + 9||\phi''||^2_{\infty})$. For each $k \geq 1$, denote the penalized empirical ranking performance measure by

$$\widehat{W}^{\phi,k}_{n,m}(\hat{s}_k)/n = \widehat{W}^{\phi}_{n,m}(\hat{s}_k)/n - B_1\sqrt{\frac{\mathscr{V}_k}{pN}} - \sqrt{\frac{2C\log k}{p^2N}} \ . \tag{5.6.52}$$

For any $\varepsilon > 0$, we have, as soon as $pN \geq B_2 \sup_{k\geq 1} \mathscr{V}_k$,

$$\begin{aligned}
\mathbb{P}\left\{ \widehat{W}^{\phi,\hat{k}}_{n,m}(\hat{s}_{\hat{k}})/n - W_{\phi}(\hat{s}_{\hat{k}}) \geq \varepsilon \right\} &\leq \sum_{k\geq 1} \mathbb{P}\left\{ \widehat{W}^{\phi,k}_{n,m}(\hat{s}_k)/n - W_{\phi}(\hat{s}_k) \geq \varepsilon \right\} \\
&\leq \sum_{k\geq 1} \mathbb{P}\left\{ \widehat{W}^{\phi}_{n,m}(\hat{s}_k)/n - B_1\sqrt{\frac{\mathscr{V}_k}{pN}} - W_{\phi}(\hat{s}_k) > \varepsilon + \sqrt{\frac{2C\log k}{p^2N}} \right\} \\
&\leq \sum_{k\geq 1} \exp\left(-\frac{p^2N}{C}\left(\varepsilon + \sqrt{\frac{2C\log k}{p^2N}} \right)^2 \right) \\
&\leq \exp\left(-\frac{p^2N\varepsilon^2}{C} \right) \sum_{k\geq 1} k^{-2} < 2\exp\left\{ -\frac{p^2N\varepsilon^2}{C} \right\} \ . \qquad (5.6.53)
\end{aligned}$$

For all $k \geq 1$, $W_k^* = \sup_{s \in \mathscr{S}_k} W_\phi(s) = W_\phi(s_k^*)$ and consider the decomposition

$$W_k^* - W_\phi(\hat{s}_{\hat{k}}) = \left(W_k^* - \widehat{W}_{n,m}^{\phi,\hat{k}}(\hat{s}_{\hat{k}})/n\right) + \left(\widehat{W}_{n,m}^{\phi,\hat{k}}(\hat{s}_{\hat{k}})/n - W_\phi(\hat{s}_{\hat{k}})\right) .$$

The expectation of the second term of the right hand side of the equation above can be bounded by means of the tail bound (5.6.53)

$$\mathbb{E}\left[\widehat{W}_{n,m}^{\phi,\hat{k}}(\hat{s}_{\hat{k}})/n - W_\phi(\hat{s}_{\hat{k}})\right] \leq 2\sqrt{\frac{C}{p^2 N}} . \tag{5.6.54}$$

for any $k \geq 1$, as soon as $pN \geq B_2 \sup_{k \geq 1} \mathscr{V}_k$. Concerning the expectation of the first term, observe that

$$\begin{aligned}\mathbb{E}\left[W_k^* - \widehat{W}_{n,m}^{\phi,\hat{k}}(\hat{s}_{\hat{k}})/n\right] &\leq \mathbb{E}\left[W_k^* - \widehat{W}_{n,m}^{\phi,k}(s_k^*)\right] \\ &\leq \mathbb{E}\left[W_\phi(s_k^*) - \widehat{W}_{n,m}^{\phi}(s_k^*)\right] + \mathrm{pen}(N,k) \leq B_1\sqrt{\frac{\mathscr{V}_k}{pN}} + \mathrm{pen}(N,k) ,\end{aligned}$$

for any $k \geq 1$, as soon as $pN \geq B_2 \sup_{k \geq 1} \mathscr{V}_k$. Summing the bound obtained and that in (5.6.54) gives the desired result.

5.6.6 Proof of Proposition 58

The proof consists in combining the two results stated below with the decomposition (5.4.11) of the W_ϕ-ranking performance deficit of the maximizer. The first result is the analogue of Theorem 54 for the smoothed criterion.

Theorem 66. *Suppose that the assumptions of Proposition 53 are fulfilled. Then, for any $\delta \in (0,1)$, there exist constants $C_1, C_3 > 0$, $C_2 \geq 24$, depending on ϕ, K, R, $\mathscr{V}$ such that with probability larger than $1-\delta$:*

$$\sup_{s \in \mathscr{S}_0} \left|\widehat{W}_{n,m,h}^{\phi}(s)/n - \widetilde{W}_{\phi,h}(s)\right| \leq C_3\sqrt{\frac{\log(C_2/\delta)}{pN}} , \tag{5.6.55}$$

as soon as $N \geq c/(p\min(p,1-p)^2)\log(C_2/\delta)$ and $\delta \leq C_2 e^{-(C_1/C_3)^2}$ with $c > 0$ depending on ϕ, K, R, $\mathscr{V}$.

The proof being quite similar to that of Theorem 54, it is omitted. Assumption 7 ensuring that the class $\{K((\cdot - t)/h);, t \in \mathbb{R}^q, h > 0\}$ ($q = 1$ here) is bounded VC-type (see *e.g.* Lemma 22(ii) in Nolan and Pollard (1987) and Giné et al. (2004)), classic permanence properties can be used to check that all the classes of functions over which uniform bounds are taken are of finite VC dimension. The second result provides a uniform bound for the additional bias error made when approximating $W_\phi(s)$ by $\widetilde{W}_{\phi,h}(s)$ for $s \in \mathscr{S}_0$.

Lemma 67. *Suppose that Assumptions 6 is satisfied. Then, for all $h > 0$, we have:*

$$\sup_{s \in \mathscr{S}_0} \left|\widetilde{W}_{\phi,h}(s) - W_\phi(s)\right| \leq C_4 h^2, \tag{5.6.56}$$

where $C_4 > 0$ is a constant depending on ϕ, K and R only.

Details are left to the reader, the proof is straightforward under Assumption 6, using the regularity of the score generating function and the uniform integrated error bound obtained in Jones (1990).

6 | Two-sample Homogeneity Testing

Abstract. In the continuity of Chapter 5, we apply the multivariate generalization of *R*-statistics to the two-sample problem. This leads to a generic two-stage test procedure for which theoretical guarantees are proved. The approach is composed of: (i) *Bipartite ranking:* a bipartite ranking algorithm learns the optimal scoring function in the sense of Chap. 5 on the first half of each sample, (ii) *Univariate two-sample homogeneity test:* the chosen rank test is performed on the remaining univariate observations for a given testing level, mapped with the optimal function of step (i). In fact, under regularity conditions ensured by the results of Chapter 5, (i) aims to learn strictly monotonous transforms of the likelihood ratio between the multivariate distribution functions, ignoring the curse of dimensionality and possible sampling bias issues while satisfying the rank-related properties. Regarding (ii), we establish (non)asymptotic guarantees for the bias of the *R*-statistics under the null and the alternative distributions when valued at the solution of the first step. The proposed class of statistics is shown to be distribution-free insofar it inherits from important univariate properties. We also present a procedure for choosing the optimal score-generating function in a minimax testing sense. Chapter 7 gathers the related numerical experiments.

Contents

6.1 Introduction

The nonparametric statistical hypothesis testing problem referred to as the *two-sample problem* is of central importance in statistics and machine-learning, as defined in Chapter 2, Section 2.1. Easy to formulate, this generic problem is ubiquitous. It finds applications in many areas, in particular in clinical trials, in order to determine whether the fluctuations of a collection of measurements performed over two statistical populations subject to different treatments are simply due to the sampling phenomenon or else to the effect of the treatment. It may also be used in the context of multimodal data fusion to decide whether two datasets can be pooled or not. Indeed, selection/sampling bias issues are also a major concern in machine learning now. As recently highlighted by theoretical and empirical works (see *e.g.* Bertail et al. (2021) or Wang et al. (2019)), a poor control on the acquisition process of training data, even massive, can significantly jeopardize the generalization ability of the learned predictive rules. Various dedicated approaches have been introduced in the statistical literature, see *e.g.* section 6.9 in Lehmann and Romano (2005) or Chapter 3.7 in van der Vaart and Wellner (1996). Many of them consist in computing first nonparametric estimators $\widehat{G}_n$ and $\widehat{H}_m$ of the underlying distributions (possibly smoothed versions). Next, by evaluating a distance or information-theoretic pseudo-metric $\mathscr{D}(\widehat{G}_n, \widehat{H}_m)$ between the latter in order to measure their dissimilarity (*e.g.* the two-sample Kolmogorov-Smirnov statistic), the null hypothesis is rejected for 'large' values of the statistic $\mathscr{D}(\widehat{G}_n, \widehat{H}_m)$, see Biau and Gyorfi (2005) or Gretton et al. (2012a) for instance. Beyond computational difficulties and the necessity of identifying a proper standardization in order to make the statistic asymptotically pivotal, *i.e.* its limit distribution is parameter free (this generally requires in practice the use of resampling/bootstrap techniques), the major issue one faces when trying to implement such *plug-in* procedures is related to the curse of dimensionality. Indeed, such plug-in procedures involve the consistent estimation of distributions on a feature space of possibly very large dimension $d \in \mathbb{N}^*$. We refer to the Section 2.1 for an account of state-of-the-art methods are their properties.

The approach promoted and analyzed in this chapter is very different and is inspired from *rank tests* in the univariate case, see Appendix section B.1. The rationale behind the use of two-sample rank statistics for univariate two-sample problems, naturally lies in the fact that they are distribution-free under the null hypothesis and are unbiased. Also, for an appropriate choice of ϕ, the related rank test is know to have asymptotic power with optimal properties, describing its capacity to detect alternatives close to the null assumption, see *e.g.* Hájek and Sidák (1967) or Chapter 15 in van der Vaart (1998). For instance, the popular Mann-Withney-Wilcoxon 'ranksum' statistic, widely used to test whether a distribution is stochastically larger than another one, corresponds to the case $\phi(u) = u$ and is optimal to detect asymptotically small shifts. The methodology we propose here to extend these techniques to the multivariate setup starts from the observation that, in the univariate case, two-sample rank statistics are summaries of the statistical version of the ROC curve relative to the pair of probability measures $(H,\ G)$. Under $\mathscr{H}_0$, the ROC curve coincides with the diagonal of the unit square $[0,1]^2$. In particular, up to an affine transform, the ranksum statistic is nothing else than the AUC. The method uses the 2-split trick and is implemented in two steps. First, a bipartite ranking algorithm is performed on the first half of the samples, in order to learn the optimal scoring function $s : \mathscr{X} \to \mathbb{R}$ as to rank multivariate observations. Next, a univariate rank test is applied to the other half observations that are mapped thanks to the optimal scoring function. Incidentally, we will interpret the homogeneity step as testing whether the empirical ROC curve obtained significantly deviates from the diagonal, the nature of the deviation being determined by the score generating function ϕ used in the rank test. By means of concentration results for two-sample linear rank processes (Chap. 5), the two-stage testing procedure is analyzed from a nonasymptotic perspective. We prove bounds for the bias of tests based on a scoring function s maximizing a statistical counterpart of a bipartite ranking performance criterion, taking the form of a two-sample linear rank statistic. Also, a procedure for choosing the optimal score-generating function in a minimax testing sense is proposed and the asymptotic distributions of the (studentized) statistics are proved.

Lastly, all numerical experiments are gathered in the next Chapter 7. In particular, the capacity of the two-stage method proposed to detect 'small' deviations from the null assumption, preserved even in very high dimension to a certain extent, is investigated from an empirical angle. An extensive experimental study is presented, comparing the performance of the *ranking-based tests* to that of alternative nonparametric methods documented in the literature. Notice finally that a very preliminary version of the two-stage testing method, limited to AUC optimization and 'ranksum' test statistics, has been previously outlined in the conference paper Clémençon et al. (2009).

The chapter is organized as follows. In section 6.2, the main notations are set out, the statistical framework of the two-sample problem is recalled, rank tests in the univariate case and their relation to bipartite ranking are briefly reviewed. The novel testing procedure is described in section 6.3.1. Section 6.4 details the theoretical guarantees. We consider the framework and notation of Chapter 5.

6.2 Background and preliminaries

This section introduces the main notation and the nonparametric two-sample problem. The existing methods are briefly reviewed, with particular attention to *rank tests* in the $1-d$ case and their interpretation through ROC analysis. Concepts and results related to the bipartite ranking task, viewed as the problem of optimizing (summaries of) the ROC curve, are also briefly recalled, insofar as the methodology proposed and analyzed in the subsequent section is based on the latter. We refer to Chapter 2 for detailed reviews on state-of-the-art methods.

6.2.1 The two-sample problem

Let $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$ be independent *i.i.d.* samples drawn from probability distributions G and H on a measurable space $\mathscr{X}$. In the most general version of the two-sample problem, one makes no assumption about the distributions H and G and the goal pursued is to test the composite hypothesis:

$$\mathscr{H}_0 : H = G \text{ against the alternative } \mathscr{H}_1 : H \neq G \,, \tag{6.2.1}$$

based on the two samples. As detailed in Chapter 2, section 2.1, state-of-the-art approaches are typically based on estimating a (pseudo)-metric, based on the two empirical measures or related, see Eq. (2.1.2). Beyond computational difficulties and the necessity of identifying a proper standardization to make asymptotically pivotal (*i.e.* its limit distribution is parameter free, see MMD test statistic Eq. (2.1.9)) and determining an appropriate critical threshold (this generally requires in practice the use of bootstrap techniques), the major issue one faces when trying to implement such *plug-in* procedures is related to the curse of dimensionality. Indeed, such plug-in procedures involve the consistent estimation of distributions on a feature space of possibly very large dimension $d \in \mathbb{N}^*$. This difficulty can however be circumvented to a certain extent when a unit ball of a reproducing kernel Hilbert space $\mathscr{H}$ is chosen for $\mathscr{F}$ in order to allow for efficient computation of the MMD supremum, see Gretton et al. (2007) and Bach et al. (2008). The methodology promoted in the present chapter for the two-sample problem is very different in nature and is inspired from traditional techniques in the particular one-dimensional case.

6.2.2 The univariate case - rank tests and ROC analysis

A classic approach to the two-sample problem in the one-dimensional setup consists in ranking the observed data using the natural order on the real line, and taking the decision depending on the ranks of the positive instances among the pooled sample:

$$\forall i \in \{1, \ldots, n\}, \ \mathrm{Rank}(X_i) = N\widehat{F}_N(X_i) \,,$$

where $\widehat{F}_N(t) = (1/N)(\sum_{i=1}^n \mathbb{I}\{X_i \leq t\} + \sum_{j=1}^m \mathbb{I}\{Y_j \leq t\})$ and $N = n+m$. Assuming that the distributions G and H are continuous for simplicity (the probability that ties occur is then equal to zero), the idea underlying rank tests lies in the simple fact that, under the null hypothesis $\mathscr{H}_0$, the ranks of positive instances are distribution-free, uniformly distributed over $\{1, \ldots, N\}$. A popular choice is to consider the sum of 'positive ranks', leading to the well-known *ranksum* Mann-Withney-Wilcoxon statistic, see Wilcoxon (1945):

$$\widehat{W}_{n,m} = \sum_{i=1}^{n} \mathrm{Rank}(X_i) \,. \tag{6.2.2}$$

It is widely used to test $\mathscr{H}_0$ against the alternative stipulating that one of the two distributions is stochastically larger than the other one. In the situation where $G(dx)$ is stochastically larger[1] than $H(dy)$, *i.e.* when $H(z) \geq G(z)$ for all $z \in \mathbb{R}$, the test statistic (6.2.2) is naturally expected to take 'large' values. Tables for the distribution of the statistics $\widehat{W}_{n,m}$ under $\mathscr{H}_0$ being available (even in the case where some observations are tied, by assigning the mean rank to ties, see Cheung and Klotz (1997)), no asymptotic approximation result is thus needed for building a test at an appropriate appropriate level. In the case where the two *c.d.f.* are linked by the relationship $G(x) = H(x-\theta)$ with $\theta \geq 0$

[1] Given two distribution functions $H(dt)$ and $G(dt)$ on $\mathbb{R} \cup \{+\infty\}$, it is said that $G(dt)$ is *stochastically larger* than $H(dt)$ if and only if for any $t \in \mathbb{R}$, we have $G(t) \leq H(t)$. We then write: $H \leq_{sto} G$. Classically, a necessary and sufficient condition for G to be stochastically larger than H is the existence of a coupling (X, Y) of (G, H), *i.e.* a pair of random variables defined on the same probability space with first and second marginals equal to H and G respectively, such that $X \leq Y$ with probability one.

(corresponding to the situation where the treatment effect is modeled in an additive fashion and the null assumption reduces to $\mathscr{H}_0 : \theta = 0$), the test statistic (6.2.2) is asymptotically uniformly most powerful in the limit experiment $\theta \searrow 0$, see *e.g.* Corollary 15.14 in section 15.5 of van der Vaart (1998). Other functionals of the 'positive ranks' can be used as test statistics for the two-sample problem. In particular, the class of two-sample linear rank statistics defined below forms a rich collection of functionals.

Definition 68. (TWO-SAMPLE LINEAR RANK STATISTICS) *Let $\phi : [0,1] \to [0,1]$ be a nondecreasing function. The two-sample linear rank statistics with 'score-generating function' $\phi(u)$ based on the random samples $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_m\}$ is given by*

$$\widehat{W}_{n,m}^{\phi} = \sum_{i=1}^{n} \phi\left(\frac{\mathrm{Rank}(X_i)}{N+1}\right) . \tag{6.2.3}$$

For $\phi(u) = u$, the statistic (6.2.3) coincides with $\widehat{W}_{n,m}/(N+1)$. Under $\mathscr{H}_0$, the statistics (6.2.3) defined above are all distribution-free, which make them particularly useful to detect differences between the distributions H and G. Tabulating their distribution under the null assumption, they can be used to design unbiased tests at certain levels α in $(0,1)$. The choice of the score-generating function ϕ can be guided by the type of difference between the two distributions (*e.g.* in scale, in location) one possibly expects, and may then leads to locally most powerful testing procedures, capable of detecting certain types of 'small' deviations from $\mathscr{H}_0$. Refer to Chapter 9 in Serfling (1980) or to Chapter 13 in van der Vaart (1998) for an account of the asymptotic theory of rank statistics. We also emphasize that concentration properties of two-sample linear rank processes have recently been studied in Clémençon et al. (2021) (Chap. 5), motivated by the interpretation of (6.2.3) as a scalar statistical summary of the ROC curve relative to the pair (H,G).

Relation to ROC **analysis.** As outlined in Chapter 5, two-sample linear rank statistics as defined in (6.2.3) are intimately related to univariate ROC analysis. Briefly, we recall that the AUC of the empirical ROC curve is proportional (up to an affine transform) to the *ranksum* Mann-Whitney-Wilcoxon statistic (6.2.2)

$$\widehat{W}_{n,m} = nm\mathrm{AUC}_{\widehat{\mathrm{H}}_\mathrm{m},\widehat{\mathrm{G}}_\mathrm{n}} + \frac{\mathrm{n(n+1)}}{2} .$$

More generally, two-sample linear rank statistics (6.2.3) related to score generating functions different from $\phi(u) = u$ provide alternative summaries of the empirical ROC curve and measure different ways of deviating from the main diagonal of the unit square, which coincides with $\mathrm{ROC}_{\mathrm{H,G}}$ under the null assumption. Notice incidentally that, under $\mathscr{H}_0$, we have

$$\mathrm{AUC}_{\mathrm{H,G}} - 1/2 = \int_{\alpha=0}^{1} \{\mathrm{ROC}_{\mathrm{H,G}}(\alpha) - \alpha\}\,\mathrm{d}\alpha = 0 .$$

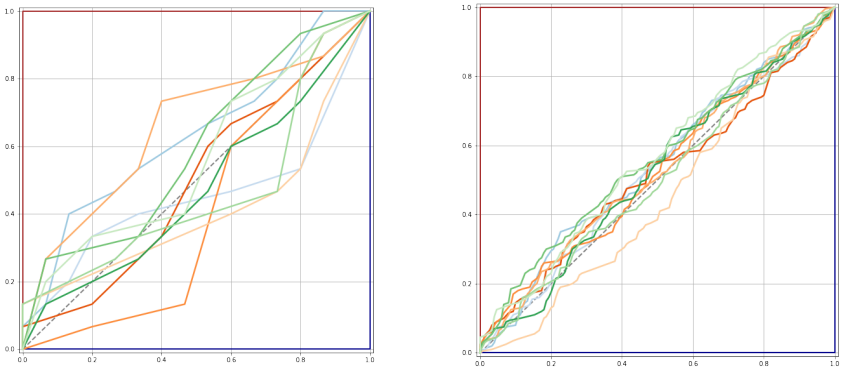
Like (6.2.2), the statistics (6.2.3) are pivotal and in order to quantify their fluctuations as the full sample sizes n, m increase, the fraction of 'positive'/'negative' observations in the pooled dataset must be controlled. Let $p \in (0,1)$ be the 'theoretical' fraction of positive instances. For $N \geq 1/p$, we suppose that $n = \lfloor pN \rfloor$ and $m = \lceil (1-p)N \rceil = N - n$. Define the mixture probability distribution $F = pG + (1-p)H$. As N tends to infinity, the asymptotic mean of $\widehat{W}_{n,m}^{\phi}/n$ is:

$$W_\phi = \mathbb{E}[\phi \circ F(X)] = \frac{1}{p}\int_0^1 \phi(u)du - \frac{1-p}{p}\int_0^1 \phi\left(p(1-\mathrm{ROC}_{\mathrm{H,G}}(\alpha)) + (1-\mathrm{p})(1-\alpha)\right)\, d\alpha \ , \tag{6.2.4}$$

see section 5.3 of Chap. 5 (section 3 in Clémençon et al. (2021)). Observe that under $\mathscr{H}_0$, we have

$$W_\phi = \int_0^1 \phi(u)du \, .$$

Building a two-sample test in the ROC space. As previously highlighted, under $\mathscr{H}_0$, the theoretical ROC curve coincides with the main diagonal of $[0,1]^2$: $\mathrm{ROC}_{H,H}(\alpha) = \alpha$ for all $\alpha \in (0,1)$ and any distribution H on $\mathbb{R}$. In addition, since the empirical ROC curve is itself a function of the ranks, it is also a (functional) pivotal statistic under the null assumption. When the probability distribution H is continuous, all the possible empirical ROC curves are equiprobable under the null assumption. Hence, in the situation where $\binom{N}{n}$ is not too large, since the ensemble $\mathscr{C}_{n,m}$ of all possible empirical ROC curves based on positive and negative samples of respective sizes n and m is of cardinality $\binom{N}{n}$, all broken lines included in it can be enumerated (see Fig. 6.1) and for any $\alpha \in \left\{ i/\binom{N}{n} : i = 1, \ldots, \binom{N}{n} \right\}$, one can build a tolerance/prediction region $\mathscr{R}_\alpha \subset \mathscr{C}_{n,m}$ of level α, *i.e.* a subset $\mathscr{R}_\alpha \subset \mathscr{C}_{n,m}$ of cardinality $\alpha\binom{N}{n}$. Then, a test that rejects $\mathscr{H}_0$ when the empirical ROC curve $\widehat{\mathrm{ROC}}_{H,G}$ falls outside $\mathscr{R}_\alpha$ can be considered. A natural way of building a critical region in the space $\mathbb{D}([0,1])$ of càd-làg mappings from $[0,1]$ to $[0,1]$ that defines a test of hypothesis $\mathscr{H}_0$ at level α is to fix a pseudo-distance D on $\mathbb{D}([0,1])$, sort the $\binom{N}{n}$ curves in $\mathscr{C}_{n,m}$ by increasing distance to the first diagonal and keep the subset $\mathscr{R}_\alpha$ formed by the $\lceil \binom{N}{n}(1-\alpha) \rceil$ curves closest to the diagonal in the sense of the distance D chosen. When choosing the distance defined by L_1-norm, one naturally recovers the Mann-Whitney-Wilcoxon test. However, many functional distances can be considered for this purpose.



a. Breakpoints of empirical ROC curves with $n, m = 15$

b. Empirical ROC curves with $n, m = 200, 150$

Figure 6.1. Examples of empirical ROC curves simulations under the null hypothesis.

Remark 7. (THE AUC AS A STATISTICAL DISTANCE) *One may easily show that*

$$\mathrm{AUC}_{H,G} = \frac{1}{2} + \int_{-\infty}^{\infty} \{H(t) - G(t)\} dH(t) \, , \tag{6.2.5}$$

see the Appendix section for further details. Hence, when H is stochastically smaller than G, the quantity $\mathrm{AUC}_{H,G} - 1/2$ is equal to the $L_1(H)$-distance between the cumulative distribution functions $H(t)$ and $G(t)$.

Extensions of rank-based tests to the multivariate framework. Given the absence of any 'natural order' on $\mathbb{R}^d$ as soon as $d \geq 2$, numerous methods have been explored to circumvent it *via*

new definitions of ranks. As exposed in the general Introduction, section 1.3, multiple types of rank concepts exist. They usually rely on depth-based Beirlant et al. (2020); Chaudhuri (1996); Chernozhukov et al. (2017); Deb and Sen (2019); Koshevoy and Mosler (1997); Liu (1990, 1995); Oja (1983); Vardi and Zhang (2000) or spatial ranks Möttönen et al. (1997, 2005); Möttönen and Oja (1995) approaches, while others are obtained *via* distance-based ranks Hallin and Paindaveine (2002a,b, 2008) . Nevertheless, the majority is derived in semiparametric frameworks due to the inherent complexity of nonparametric models.

In the next section 6.3, we detail and analyze the proposed approach. We raise that it shares similarities with the method relying on statistical depth, except that the mapping used to 'project' the multivariate observations onto the real line is specifically learned from the data in order to detect best the deviations in distribution between the two distributions. In this sense, the subsequent subsection explains how any scoring function $s : \mathscr{Z} \to]-\infty, +\infty]$ solution of the bipartite ranking problem related to the pair (H, G) permits to extend the use of ROC analysis and two-sample rank statistics to the two-sample problem in the multivariate setup developed.

6.2.3 Bipartite ranking - the rationale behind our approach

The goal of bipartite ranking is to learn, based on the 'positive' and 'negative' samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$, how to score any new observations $\mathbf{Z}_1, \ldots, \mathbf{Z}_k$, being each either 'positive' or else 'negative', that is to say drawn either from G or else from H, without prior knowledge, so that positive instances are mostly at the top of the resulting list with large probability. As detailed in Sections 2.2 and 5.3, a natural way of defining a total preorder[2] on $\mathscr{Z}$ is to map it with the natural order on $\mathbb{R} \cup \{+\infty\}$ by means of a *scoring rule*. It is defined as a measurable mapping $s : \mathscr{Z} \to]-\infty, +\infty]$ such that a preorder $\preccurlyeq_s$ on $\mathscr{Z}$ is: for all x, x' in $\mathscr{Z}$, $x \preccurlyeq_s x'$ *iff* $s(x) \leq s(x')$. We denote by $\mathscr{S}$ the set of all scoring functions. The capacity of a candidate $s(z)$ in $\mathscr{S}$ to discriminate between the positive and negative statistical populations is generally evaluated by means of the ROC curve $\mathrm{ROC}_{\mathrm{H_s,G_s}}(\alpha) = \mathrm{ROC}(\mathrm{s}, \alpha)$, denoting by H_s and G_s the pushforward probability distributions of H and G by the mapping $s(z)$. Precisely, it is shown in Clémençon and Vayatis (2009b) (see Proposition 4 therein) that the optimal scoring rules are the elements of the set

$$\mathscr{S}^* = \left\{ s \in \mathscr{S},\ \forall (z, z') \in \mathscr{Z}^2,\ \Psi(z) < \Psi(z') \Rightarrow s^*(z) < s^*(z') \right\}. \tag{6.2.6}$$

where $\Psi(z) = dG/dH(z)$ is the likelihood ration. This leads to: $\forall (s, \alpha) \in \mathscr{S} \times (0,1)$,

$$\mathrm{ROC}(\mathrm{s}, \alpha) \leq \mathrm{ROC}^*(\alpha) ,$$

where $\mathrm{ROC}^*(.) = \mathrm{ROC}(\Psi, .) = \mathrm{ROC}(\mathrm{s}^*, .)$ for any s^* in $\mathscr{S}$. Recall incidentally that this optimal curve is non-decreasing and concave and thus always above the main diagonal of the unit square. The bipartite ranking formulations *via* ROC analysis are extensively detailed in Section 2.2 of Chap. 5.

Ranking-based two-sample rank tests. The two-sample test procedure 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 $\int_0^1 \phi(u)du$ for appropriate score generating functions ϕ, provide a natural way of measuring the dissimilarity beween G and H in theory. As revealed by the proposition below, such deviations are equal to zero as soon as the null assumption is fulfilled.

Proposition 69. *The following assertions are equivalent.*

[2] A preorder $\preccurlyeq$ on a set $\mathscr{Z}$ is a reflexive and transitive binary relation on $\mathscr{Z}$. It is said to be *total*, when either $z \preccurlyeq z'$ or else $z' \preccurlyeq z$ holds true, for all $(z, z') \in \mathscr{Z}^2$.

(i) *The assumption '$\mathscr{H}_0 : H = G$' holds true*

(ii) *The optimal* ROC *curve relative to the bipartite ranking problem defined by the pair (H,G) coincides with the diagonal of $[0,1]^2$:*

$$\forall \alpha \in (0,1),\ \mathrm{ROC}^*(\alpha) = \alpha\ .$$

(iii) *For any score generating function $\phi(u)$, we have:*

$$W_\phi^* = \int_0^1 \phi(u)du\ .$$

(iv) *There exists a strictly increasing score generating function $\phi(u)$, such that:*

$$W_\phi^* = \int_0^1 \phi(u)du\ .$$

(iv) *We have:* $\mathrm{AUC}^* = 1/2$.

In addition, we have:

$$\mathrm{AUC}^* - 1/2 = \mathbb{E}[|\Psi(\mathbf{Y}) - 1|]\ . \tag{6.2.7}$$

We also recall that the optimal ROC curve related to the pair of distributions (H,G) is the same as that related to the pair of univariate distributions (H_{s^*}, G_{s^*}) and that $dG/dH(z) = dG_{s^*}/dH_{s^*}(s^*(z))$ for any $s^* \in \mathscr{S}^*$, see Corollary 5 in Clémençon and Vayatis (2009b). Hence, the optimal curve ROC^* is a very natural and exhaustive way of measuring the dissimilarity between two multivariate distributions, extending the basic ROC analysis for distributions on $\mathbb{R}$ recalled in subsection 6.2.2, as illustrated by the example below.

Example 70. (MULTIVARIATE GAUSSIAN POPULATIONS) *Consider two Gaussian distributions H and G on $\mathbb{R}^d$ with same positive definite covariance matrix Γ and respective means θ_- and θ_+ in $\mathbb{R}^d$, supposed to be distinct. As an increasing transform of the* log*likelihood ratio, the scoring function:*

$$s(z) = \langle z,\ \Gamma^{-1}(\theta_+ - \theta_-)\rangle,\ \ z \in \mathbb{R}^d\ ,$$

is optimal scoring function, denoting by $\langle .,\ .\rangle$ the usual Euclidean inner product on $\mathbb{R}^d$. Since it is linear, the distributions H_s and G_s are both Gaussian univariate distributions. Denoting $\Delta(t) = (1/\sqrt{2\pi})\int_{-\infty}^t \exp(-u^2/2)du$, $t \in \mathbb{R}$, the cdf of the centered standard univariate Gaussian distribution, one may immediately check that the optimal ROC *curve is given by:*

$$\forall \alpha \in (0,1),\ \mathrm{ROC}^*(\alpha) = 1 - \Delta\left(\Delta^{-1}(1-\alpha) - \sqrt{\mathrm{s}(\theta_+ - \theta_-)}\right).$$

And, the optimal AUC *yields to:*

$$\mathrm{AUC}^* - 1/2 = 1 - \exp\{\mathrm{s}(\theta_+ - \theta_-)\}\ .$$

If now $\theta_+ = \theta_- = 0_d$, $\Gamma_+ = I_d$ and Γ_- is symmetric positive definite, it exists therefore a orthonormal basis that diagonalizes Γ_-. Consider u_M the eigenvector associated to the highest eigenvalue of the covariance matrix $\lambda_M(\Gamma_-)$. An increasing transform of the log*likelihood ratio can be obtained by the scoring function $s(z) = \langle z,\ u_M\rangle$, $z \in \mathbb{R}^d$. It follows that $\forall \alpha \in (0,1)$, $\mathrm{ROC}^*(\alpha) = 1 - \Delta\left(\lambda_{\mathrm{M}}(\Gamma_-)^{1/2}\Delta^{-1}(1-\alpha)\right)$.*

More generally, if the two distributions are of the group of multivariate α-stable distribution, thanks to the linear stability property, projecting observations has no effect on the type of distribution. It is straightforward to express the likelihood ratio when the class of scoring functions is considered as linear i.e. $\mathscr{S}_\Theta := \{s_\theta(\cdot) := \langle \cdot,\ \theta\rangle,\ \theta \in \Theta\}$. This property will be used for the numerical examples and will be discussed in the Appendix.

As the optimal ROC curve and its summaries such as the quantities AUC^* or W_ϕ^* are unknown in practice, the approach we propose for solving the two-sample problem is implemented in two steps, splitting the samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$ into two halves: 1) the first step consists in solving the bipartite ranking problem based on the first halves of the 'positive' and 'negative' samples, as described in the preceding subsection, producing a scoring function $\widehat{s}(z)$, 2) in the second step of the procedure, the scores of the remaining data are first computed by means of the function $\widehat{s}(z)$ learned at the previous step and a rank-based test is next performed based on the latter. The subsequent sections of the present chapter provide both theoretical and empirical evidence that, beyond the fact that they are nearly unbiased, such testing procedures permit to detect very small deviations from the null assumption.

6.3 Ranking-based rank tests for the two-sample problem

This section describes at length the two-sample methodology foreshadowed by the observations made in the previous section and discuss its possible implementations from a practical perspective.

6.3.1 Method

We now explain how the general idea sketched in subsection 6.2.3 can be applied effectively, based on the observation of two independent *i.i.d.* samples $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ with $n, m \geq 1$. Let $\alpha \in (0,1)$ be the target level, *i.e.* the desired type-I error. As previously discussed, two ingredients are essentially involved in the testing procedure:

1. A bipartite ranking algorithm $\mathscr{A} : \mathscr{Z}^{n+m} \to \mathscr{S}_0$ operating on a class $\mathscr{S}_0 \subset \mathscr{S}$ and assigning to any set of training observations $\mathscr{D}_{n,m} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \cup \{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$ a scoring function $\mathscr{A}(\mathscr{D}_{n,m})$ in $\mathscr{S}_0$;

2. A two-sample rank test $\Phi_\alpha^\phi :]-\infty, +\infty]^{n+m} \to \{0, 1\}$ of level $\alpha \in (0,1)$ with outcome depending on $\{\mathrm{Rank}(x_1), \ldots, \mathrm{Rank}(x_n)\}$ for any pooled univariate dataset $\mathscr{D}_{n,m} = \{x_1, \ldots, x_n\} \cup \{y_1, \ldots, y_m\}$ and score-generating function ϕ.

Equipped with these two components, the methodology is implemented in two main steps, as detailed in Figure 6.2.

Remark 8. (ON BIPARTITE RANKING ALGORITHMS) *As mentioned in subsection 6.2.3, the vast majority of bipartite ranking algorithms documented in the statistical learning literature solve M-estimation problems over specific classes $\mathscr{S}_0$ of scoring functions. The criterion one seeks to maximize is the* AUC *or a (smoothed/concavified/penalized) variant of the latter such as (6.2.4), whose set of optimal elements coincide with a subset of $\mathscr{S}^*$. See e.g. Freund et al. (2003), Rakotomamonjy (2004), Rudin et al. (2005), Rudin (2006) or Burges et al. (2007) Generalization results in the form of confidence upper bounds for the deficit of empirical maximizers have been established under various complexity assumptions for $\mathscr{S}_0$ in Clémençon et al. (2008), Agarwal et al. (2005), Clémençon and Vayatis (2007), Clémençon et al. (2011), Menon and Williamson (2016) and Clémençon et al. (2021). Stronger theoretical guarantees (i.e. bounds for the* sup*-norm deviation (2.2.7)) have also been established for alternative approaches, considering* ROC *optimization as a continuum of cost-sensitive binary classification problems and combining M-estimation with nonlinear approximation methods, see Clémençon and Vayatis (2009b), Clémençon and Vayatis (2010), Clémençon and N.Vayatis (2009) or Clémençon et al. (2013a).*

Remark 9. (ON THE TWO-SPLIT TRICK) *As recalled above, nearly optimal scoring functions are generally learned by means of M-estimation techniques. Consequently, their dependence on the*

RANKING-BASED TWO-SAMPLE RANK TESTS

Input. Two independent and *i.i.d.* samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$ of sizes n, $m \geq 2$ and valued in $\mathscr{X}$ - subsample sizes $n' < n$ and $m' < m$ - bipartite ranking $\mathscr{A}$ algorithm operating on the class $\mathscr{S}_0$ of scoring functions on $\mathscr{X}$ - univariate two-sample rank test Φ_α of level $\alpha \in (0,1)$

Two-split trick. Divide each of the original samples into two subsamples:

$$\{\mathbf{X}_1, \ldots, \mathbf{X}_{n'}\} \cup \{\mathbf{X}_{1+n'}, \ldots, \mathbf{X}_n\} \text{ and } \{\mathbf{Y}_1, \ldots, \mathbf{Y}_{m'}\} \cup \{\mathbf{Y}_{1+m'}, \ldots, \mathbf{Y}_m\}$$

1. **Bipartite ranking.** Run bipartite ranking algorithm $\mathscr{A}$ based on training data $\mathscr{D}_{n',m'} = \{\mathbf{X}_1, \ldots, \mathbf{X}_{n'}\} \cup \{\mathbf{Y}_1, \ldots, \mathbf{Y}_{m'}\}$, producing the scoring function
$$\widehat{s}(z) = \mathscr{A}(\mathscr{D}_{n',m'}). \tag{6.3.1}$$

2. **Univariate rank test.** Form the univariate samples
$$\{\widehat{s}(\mathbf{X}_{1+n'}), \ldots, \widehat{s}(\mathbf{X}_n)\} \text{ and } \{\widehat{s}(\mathbf{Y}_{1+m'}), \ldots, \widehat{s}(\mathbf{Y}_m)\},$$
compute the α-level test statistic
$$\Phi_\alpha^\phi\left(\{\widehat{s}(\mathbf{X}_{1+n'}), \ldots, \widehat{s}(\mathbf{X}_n)\} \cup \{\widehat{s}(\mathbf{Y}_{1+m'}), \ldots, \widehat{s}(\mathbf{Y}_m)\}\right), \tag{6.3.2}$$
depending on $\{\mathrm{Rank}(\widehat{s}(\mathbf{X}_{1+n'}), \ldots, \mathrm{Rank}(\widehat{s}(\mathbf{X}_n))\}$, where $\mathrm{Rank}(t) = \sum_{i=1+n'}^{n} \mathbb{I}\{\widehat{s}(\mathbf{X}_i) \leq t\} + \sum_{i=1+m'}^{m} \mathbb{I}\{\widehat{s}(\mathbf{Y}_j) \leq t)\}$.

Figure 6.2. Ranking-based two-sample rank test procedure.

training observations may be complex and can hardly be make explicit in general. For this reason, a 2*-split trick is used in order to make the analysis of the fluctuations of the quantity* (6.3.2) *tractable. Hence, conditioned upon the subsamples used in the bipartite ranking step of the procedure, the functional* (6.3.2) *is a two-sample rank statistic.*

We now propose several ways of implementing the methodology summarized in Fig. 6.2, which will be next studied theoretically in specific situations in Section 6.4 and whose performance will be empirically investigated at length in Section 7.2, Chap. 7.

6.3.2 Ranking-based two-sample linear rank tests.

The simplest implementation consists in considering a test based on a two-sample linear rank statistic (6.2.3), characterized by a given score generating function ϕ, see Definition 68. As recalled in subsection 6.2.2, such a statistic is pivotal under $\mathscr{H}_0$, its probability distribution can be easily tabulated, even in the case where n, m take very large values, given the computing power now at disposal. For all n, $m \geq 1$ and any $\alpha \in (0,1)$, one may thus determine the quantile
$$q_{n,m}^\phi(\alpha) = \inf\left\{t \geq 0 : \mathbb{P}_{\mathscr{H}_0}\left\{\widehat{W}_{n,m}^\phi \leq t\right\} \geq 1-\alpha\right\}, \tag{6.3.3}$$

as well as the critical region

$$\left\{\widehat{W}^{\phi}_{n,m} > q^{\phi}_{n,\,m}(\alpha)\right\} \tag{6.3.4}$$

occuring with probability less than α under $\mathscr{H}_0$ in the univariate case and defining the test at level α:

$$\Phi^{\phi}_{\alpha}(\mathscr{D}_{n,m}) = \mathbb{I}\left\{\widehat{W}^{\phi}_{n,m} > q^{\phi}_{n,\,m}(\alpha)\right\}\,, \tag{6.3.5}$$

based on the univariate samples $\mathscr{D}_{n,m} = \{\{X_1, \ldots, X_n\}, \{Y_1, \ldots, Y_m\}\}$. Attention shoud be paid, that in contrast to the univariate situation, for which no bipartite ranking step is required, only a unilateral test Φ_α is relevant in the multivariate case, insofar as the two-sample testing problem (6.2.1) can be rephrased as follows:

$$\mathscr{H}_0 :\ W^*_{\phi} = \int_0^1 \phi(u)du \ \textit{versus}\ \mathscr{H}_1 :\ W^*_{\phi} > \int_0^1 \phi(u)du\,, \tag{6.3.6}$$

when ϕ is strictly increasing, see Proposition 69. As investigated in Clémençon et al. (2021), a natural bipartite ranking approach consists in maximizing a statistical version of the performance criterion (6.2.4) based on the (multivariate) training data $\mathscr{D}_{n',m'}$ over the class $\mathscr{S}_0$, *i.e.* solving the optimization problem

$$\max_{s\in\mathscr{S}_0} \widehat{W}^{\phi}_{n',m'}(s)\,, \tag{6.3.7}$$

where, for any scoring function $s(z)$, we set

$$\widehat{W}^{\phi}_{n',m'}(s) = \frac{1}{n'}\sum_{i=1}^{n'} \phi\left(\frac{\mathrm{Rank}(s(\mathbf{X}_i))}{N'+1}\right)\,, \tag{6.3.8}$$

with $N' = n' + m'$, the quantity $\mathrm{Rank}(s(\mathbf{X}_i))/(N'+1)$ being a natural empirical counterpart of $F_s(s(\mathbf{X}_i))$ for $i = 1, \ldots, n'$. The generalization capacity of solutions of the problem (6.3.7) and (gradient ascent based) optimization strategies for solving the latter approximately have been studied in Clémençon et al. (2021). Hence, provided that a solution $\widehat{s}$ of (6.3.7) has been obtained as the outcome of *Step 1*, the test built at *Step 2* based on the scored data

$$\mathscr{D}_{n'',m''}(\widehat{s}) = \{\{\widehat{s}(\mathbf{X}_{1+n'}), \ldots, \widehat{s}(\mathbf{X}_n)\}, \{\widehat{s}(\mathbf{Y}_{1+m'}), \ldots, \widehat{s}(\mathbf{Y}_m)\}\}\,, \tag{6.3.9}$$

writes

$$\Phi^{\phi}_{\alpha}(\mathscr{D}_{n'',m''}(\widehat{s})) = \mathbb{I}\left\{\widehat{W}^{\phi}_{n'',m''}(\widehat{s}) > q^{\phi}_{n'',m''}(\alpha)\right\}\,, \tag{6.3.10}$$

with $n'' = n - n'$ and $m'' = m - m'$. Under specific assumptions (related to the class $\mathscr{S}_0$, n' and m' in particular), the properties of the test (6.3.10) are investigated in subsection 6.4.1.

6.4 Theoretical guarantees

This section focuses on theoretical guarantees for the proposed two-stage procedure, in particular regarding the two-sample homogeneity test step (*Step 2* of Proc. 6.2). We consider the dataset $\mathscr{D}_{n'',m''} = \{\mathbf{X}_{1+n'}, \ldots, \mathbf{X}_n\} \cup \{\mathbf{Y}_{1+m'}, \ldots, \mathbf{Y}_m\}$, with $n'' = n - n'$ and $m'' = m - m'$, such that letting $p \in (0,1)$ the 'theoretical' fraction of the first sample and for $N'' = n'' + m'' \geq 1/p$, we suppose that $n''/N'' \to p$ and $m''/N'' \to 1-p$. From now on we will drop the primes and consider the simple indices n, m and $\mathscr{D}_{n,m}$.

Consider a fixed subclass $\mathscr{S}_0 \subset \mathscr{S}$ over which the bipartite ranking algorithm outputs the optimal scoring function $\hat{s} \in \mathscr{S}_0$. In the following, we analyze the statistic $\widehat{W}^{\phi}_{n,m}$ when valued at $\hat{s}$, *i.e.* ,

the one obtained by *Step 1*. First, for a given score-generating function $\phi(u)$, we obtain nonasymptotic concentration bounds for the statistic under fixed alternatives and under the null, as well as its (studentized) explicit asymptotic distributions. We also propose a method, for choosing the optimal function ϕ in a *minimax* sense, to control both statistical errors (type-I and type-II). To this end, we formulate the required assumptions below.

***Assumption* 8.** *Let $M > 0$. For all $s \in \mathscr{S}_0$, the random variables $s(\mathbf{X})$ and $s(\mathbf{Y})$ are continuous, with density functions that are twice differentiable and have Sobolev $\mathscr{W}^{2,\infty}$-norms bounded by $M < +\infty$.*

***Assumption* 9.** *The score-generating function $\phi : [0,1] \mapsto \mathbb{R}$, is nondecreasing and (i) twice continuously differentiable or (ii) $(1/N)\sum_{i\leq N} \phi^2(i/(N+1)) \to \int_0^1 \phi^2(u)du < \infty$, when $N \to \infty$.*

Notice that Assumptions 8 and 9(i) are the ones considered in the previous Chapter 5.

6.4.1 Concentration bounds under both testing hypothesis

This section provides exact distribution-free probability tail bounds of the deviations of the linear rank statistic $\widehat{W}_{n,m}^{\phi}$, valued at the solution $\hat{s}$ of *Step 1*. We determine its explicit confidence bounds of its distribution, when based on its continuous counterpart W_ϕ and conditionally on $\hat{s}$. We first derive a concentration bound for the two-sample linear rank statistic $\widehat{W}_{n,m}^{\phi}$ under fixed alternatives.

Proposition 71. *Suppose that Assumptions 8 and 9(i) are fulfilled. Then, based on the two samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$, for all $t > 0$, $N \geq 2$ and $s \in \mathscr{S}_0$:*

$$\mathbb{P}_{\mathscr{H}_1}\left\{ |\widehat{W}_{n,m}^{\phi}(s) - W_\phi(s)| > t \right\} \leq 18e^{-CNt^2} , \tag{6.4.1}$$

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

We highlight that for small values of N, the test statistic can be compared to the distribution of the statistic $\widehat{W}_\phi = (1/n)\sum_{i=1}^n (\phi \circ F_s)(s(\mathbf{X}_i))$, see Prop. 53. The bound would be slightly different with the constant C being proportional to $1/(\kappa_p\|\phi'\|_\infty^2)$. The proof of the Proposition is detailed in the appendix section 6.6.3. It relies on the linearization result in Prop. 53, Chapter 5 (Proposition 4 in Clémençon et al. (2021)), combined with probabilistic tail inequalities such as the one of Hoeffding (1963). In fact, the assumptions of the referenced result are here simplified as we only look at the deviations of the statistic valued at fixed scoring functions, and not uniformly over its class $\mathscr{S}_0$. Also, we address particular attention to the remainder process of the linearization: it encompasses statistics of higher order defined as degenerate U-processes.

On the contrary, the analysis is much simpler under the null hypothesis $\mathscr{H}_0$. Start by noticing for all $s \in \mathscr{S}_0$

$$\text{Rank}(s(\mathbf{X}_i)) \overset{d}{=} U_i, \text{ for all } i \leq n , \tag{6.4.2}$$

where the *i.i.d.* sequence of the *r.v.* $U_1, \ldots, U_n$ is drawn from the Uniform law on $\{1, \ldots, N\}$. The statistic is therefore independent of the scoring function and equals in distribution to $\sum_{i\leq n} \phi(U_i/(N+1))$ of mean $\widehat{W}_\phi^0 = (1/N)\sum_{i\leq N} \phi(i/(N+1))$. This simplest structure leads to the following probabilistic tail inequality, holding true for more general conditions on the score-generating function.

Proposition 72. *Suppose that Assumptions 8 and 9(ii) are fulfilled. Under the null hypothesis $\mathscr{H}_0$, the following inequality holds true for all $t > 0$ and for all $s \in \mathscr{S}_0$:*

$$\mathbb{P}_{\mathscr{H}_0}\left\{ |\widehat{W}_{n,m}^{\phi}(s) - W_\phi^0| > t \right\} \leq 2e^{-2pN(t-\Delta_N)^2} , \tag{6.4.3}$$

where $\Delta_N = |(1/N)\sum_{i\leq N} \phi(i/(N+1)) - \int_0^1 \phi(u)du|$ and $W_\phi^0 = \int_0^1 \phi$.

This result provides an upperbound of the bias of the statistic. Also, it can be interpreted as an approximation of the exact quantile of the unknown null distribution of the statistic detailed as follows. For a fixed level $\alpha \in (0,1)$, $\mathscr{H}_0$ is rejected as soon as $(\widehat{W}^{\phi}_{n,m} - W^0_{\phi})$ exceeds $z_\alpha = \pm\, \Delta_N + \sqrt{\log(1/\alpha)/(2pN)}$. We highlight the simplicity of the rejecting rule insofar it only depends on the intrinsic parameters to the model ϕ, N, p and on the level α, despite the high-dimensional setting and the unknown learning algorithm. When valued at $\hat{s}$, it is supposed to satisfy generalization properties in the bipartite ranking sense and that the related convergence is met as detailed in Chapter 5. For completeness, the section 6.4.3 focuses on the asymptotic laws of the statistic $\widehat{W}^{\phi}_{n,m}(s)$ under both hypothesis and valued at fixed scoring function s.

6.4.2 Nonasymptotic control of the testing errors

Let $\phi(u)$ a score-generating function and recall the definition of the power related to the α-level test, based on the two-sample $\mathscr{D}_{n,m}$,

$$\begin{aligned} \pi_{n,m}(\phi, \hat{s}) &= \mathbb{P}_{\mathscr{H}_1}\{\Phi^{\phi}_{\alpha}(\mathscr{D}_{n,m}(\hat{s})) = 1\} \\ &= \mathbb{P}_{\mathscr{H}_1}\{\widehat{W}^{\phi}_{n,m}(\hat{s}) > q^{\phi}_{n,m}(\alpha)\}\ , \end{aligned}$$

where $\hat{s}$ is the optimal scoring function in the sense of *Step 1* for a score-generating function ϕ and $q^{\phi}_{n,m}(\alpha)$ is the $(1-\alpha)$-quantile of the null distribution, see Eq. (6.3.4). We propose an estimator of the power function based on $M \in \mathbb{N}^*$ Monte Carlo samplings of the alternative distribution of the sample $\mathscr{D}_{n,m}$ denoted $\mathscr{D}^{(i)}_{n,m} = \mathscr{D}^{(i)}_{n',m'} \cup \mathscr{D}^{(i)}_{n'',m''}$, with $i \leq M$, defined by

$$\hat{\pi}_{n'',m''}(\phi, \hat{s}) = \frac{1 + \sum_{i \leq M} \mathbb{I}\{\widehat{W}^{\phi}_{n'',m''}(\hat{s}, \mathscr{D}^{(i)}_{n'',m''}) > q^{\phi}_{n'',m''}(\alpha)\}}{1+M}\ . \tag{6.4.4}$$

This estimator is computed for a fixed score-generating function. Nevertheless, as highlighted in Clémençon et al. (2021), tailoring the W_ϕ-criterion *w.r.t.* ϕ leads to different summaries of the ROC curve. In this sense, the subsequent paragraph outlines a test statistics aiming to enhance the testing power, by learning the 'optimal' score-generating function in the *minimax* sense.

Optimal minimax *R*-statistic. While the (uniform) control of the type-I error of the tests related to the *R*-statistics is obtained for all functions ϕ, we propose a method to choose the optimal one in the minimax sense, to obtain nonasymptotic guarantees on the control of the type-II error. Consider the class $\mathscr{P}$ of density functions satisfying the Assumption 8 for all $s \in \mathscr{S}_0$, and the class $\mathscr{C}$ of score-generating functions $\phi = [0,1] \to \mathbb{R}$, nondecreasing, satisfying Assumption 9. Denote by $\alpha \in (0,1)$ the uniform upperbound of the type-I error $\sup_{(g,h)\in\mathscr{P},\, g=h} \mathbb{P}_{\mathscr{H}_0}\{\Phi^{\phi}_{\alpha}(\mathscr{D}_{n,m}(\hat{s})) = 1\} \leq \alpha$ and by $\beta \in (0,1)$ the bound of the type-II error as $\sup_{(g,h)\in\mathscr{P},\, H \leq_{sto} G} \mathbb{P}_{\mathscr{H}_1}\{\Phi^{\phi}_{\alpha}(\mathscr{D}_{n,m}(\hat{s})) = 0\} \leq \beta$.

$$\rho(\phi, \mathscr{P}, \beta) := \inf_{\rho \in (1/2,1)} \left\{ \inf_{(g,h)\in\mathscr{P}_\rho,\, H\leq_{sto}G} \mathbb{P}_{\mathscr{H}_1}\left\{\Phi^{\phi}_{\alpha}(\mathscr{D}_{n,m}(\hat{s})) = 1\right\} \geq 1-\beta \right\} \tag{6.4.5}$$

$$= \inf_{\rho \in (1/2,1)} \left\{ \sup_{(g,h)\in\mathscr{P}_\rho,\, H\leq_{sto}G} \mathbb{P}_{\mathscr{H}_1}\left\{\Phi^{\phi}_{\alpha}(\mathscr{D}_{n,m}(\hat{s})) = 0\right\} \leq \beta \right\}\ . \tag{6.4.6}$$

The minimax rate of testing is the optimal rate obtained among all tests indexed by the class $\mathscr{C}$ that are of level α. It is given by

$$\rho(\phi^*, \mathscr{P}, \beta) = \inf_{\phi \in \mathscr{C}} \rho(\phi, \mathscr{P}, \beta)\ . \tag{6.4.7}$$

In particular, if a score-generating function exists such that $\rho(\phi, \mathscr{P}, \beta) \leq K\rho(\phi^*, \mathscr{P}, \beta)$, with $K > 0$ constant, then, the test associated to ϕ^* is said to be optimal in the minimax sense. Notice that minimax tests applied to independence testing are usually related to separation rates using L_2 distance, mutual information, *etc.* We refer to the works of Albert et al. (2021); Baraud (2002); Berrett et al. (2021); Birgé and Massart (1998); Ingster and Suslina (2000); Lepski and Spokoiny (1999); Schrab et al. (2021). Therefore, the definition above is completely new and driven by the capacity of the AUC to summarize the dissimilarity between two *d.f.* in the ROC space. As outlined in section 6.2.3, one can also refer to Clémençon et al. (2021) and Menon and Williamson (2016), which much developed it in the context of bipartite ranking. However, the definition (6.4.5) relies of the oracle AUC*, that is unknown in practice. The class of *true* optimal scoring functions $\mathscr{S}^*$ depends on the underlying distributions, see Proposition 6 in Clémençon et al. (2021) therein. Hence, if *Step 1* is independent of $\mathscr{C}$ and if consistency is obtained, we can approximate the functions of $\mathscr{S}^*$ by the solution $\hat{s}$ of the chosen bipartite ranking algorithm such that $\mathscr{P}_\rho(\hat{s}) = \{(g,h) \in \mathscr{P},\ \mathrm{AUC}_{\mathrm{h}_{\hat{s}},\mathrm{g}_{\hat{s}}} > \rho\}$.

The goal is to find the optimal score-generating function ϕ^* such that the resulting two-sample test rejects the null hypothesis with minimal uniform separation rate. Therefore the nonasymptotic control of $\rho(\phi^*, \mathscr{P}, \beta)$ boils down to the analysis of the uniform separation rate, to determine the optimal elements of $\mathscr{C}$.

6.4.3 Asymptotic guarantees

This section provides the explicit asymptotic laws of the two-sample linear rank statistic $\widehat{W}^{\phi}_{n,m}(s)$ under fixed alternatives and under the null hypothesis. The particular choice of $\hat{s}$ corresponds to Procedure 6.2. However classical, these results remain of interest for establishing classical properties such as asymptotic (relative) efficiency of the tests depending on the choice of the score-generating function $\phi(u)$, and the explicit estimation of the asymptotic quantile of the null distribution. Like the nonasymptotic bounds, we obtain exact distribution-free law under $\mathscr{H}_0$ that only depends on $\phi(u)$. The proofs are not detailed as these results exclusively rely on the central limit theorem.

Asymptotic distribution of the statistic under fixed alternatives. Two results are provided for the asymptotic distribution of the test statistic, whether it is studentized or based on its continuous counterpart. We recall the unbiased empirical counterpart of the W_ϕ functional by $\widehat{W}_\phi = (1/n)\sum_{i=1}^n (\phi \circ F_s)(s(\mathbf{X}_i))$.

Proposition 73. *Suppose that Assumptions 8 and 9(i) are fulfilled. Under the alternative hypothesis $\mathscr{H}_1$ and for all $s \in \mathscr{S}_0$, the linear R-statistic $\widehat{W}^{\phi}_{n,m}(s)$ based on the two samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$, converges in distribution to:*

$$\frac{\widehat{W}^{\phi}_{n,m}(s) - \widehat{W}_\phi(s)}{\sqrt{(N/n)\hat{\sigma}^2_\phi(s)}} \xrightarrow[N\to\infty]{d} W \sim \mathscr{N}(0,1)\ , \tag{6.4.8}$$

where considering $\mathbf{X}', \mathbf{X}$ i.i.d. drawn from $G(dt)$ and $\mathbf{Y}$ from $H(dt)$:

$$\hat{\sigma}^2_\phi(s) = \mathrm{Var}(\phi \circ F_s(s(\mathbf{X}))) + (\frac{n-1}{N+1})^2 \mathrm{Var}(k_s(\mathbf{X})) + \frac{nm}{(N+1)^2}\mathrm{Var}(\ell_s(\mathbf{Y})) + \frac{n-1}{N+1}\mathrm{Cov}(\phi \circ F_s(s(\mathbf{X}))), k_s(\mathbf{X}))\ ,$$

and

$$\begin{aligned} k_s(\mathbf{X}) &= \mathbb{E}_{s(\mathbf{X}')\sim G_s}[\mathbb{I}\{s(\mathbf{X}) \leq s(\mathbf{X}')\}\phi' \circ F_s(s(\mathbf{X}')) \mid s(\mathbf{X})]\ , \\ \ell_s(\mathbf{Y}) &= \mathbb{E}_{s(\mathbf{X})\sim G_s}[\mathbb{I}\{s(\mathbf{Y}) \leq s(\mathbf{X})\}\phi' \circ F_s(s(\mathbf{X})) \mid s(\mathbf{Y})]\ . \end{aligned}$$

Proposition 73 highlights the importance of the convergence of *Step 1* to output the optimal $\hat{s}$. This studentized statistic can be used to prevent form misuses of rank statistics due to mispecification of the null hypothesis, as detailed in Chung and Romano (2016). A straightforward limit distribution is obtained below when the test is centered *w.r.t.* its continuous counterpart $W_\phi(s)$, as $\hat{\sigma}^2_\phi(s)$ is a consistent estimator of $\sigma^2_\phi(s)$.

Proposition 74. *Suppose that Assumptions 8 and 9(i) are fulfilled. Under the alternative hypothesis $\mathscr{H}_1$ and for all $s \in \mathscr{S}_0$, the linear R-statistic $\widehat{W}^\phi_{n,m}(s)$ based on the two samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$, converges in distribution to:*

$$\sqrt{N}\left(\widehat{W}^\phi_{n,m}(s) - W_\phi(s)\right) \xrightarrow[N\to\infty]{d} W \sim \mathscr{N}(0, \sigma^2_\phi(s)) \,, \tag{6.4.9}$$

where

$$\sigma^2_\phi(s) = (1/p)\mathrm{Var}(\phi \circ F_s(s(\mathbf{X}))) + p\mathrm{Var}(k_s(\mathbf{X})) + (1-p)\mathrm{Var}(\ell_s(\mathbf{Y})) \\ + \mathrm{Cov}(\phi \circ F_s(s(\mathbf{X}))), k_s(\mathbf{X})) \,,$$

Notice that for $\phi(u) = u$, its derivative being equal to 1, the last Proposition recovers the asymptotic law of the ranksum statistic.

Asymptotic distribution of the statistic under the null hypothesis. Under the null hypothesis, the vector of univariate ranks is known to by uniformly drawn on the set of the $N!$ permutations of the integers $\{1, \ldots, N\}$ (Hájek and Sidák (1967), Lemma 13.1, van der Vaart (1998)). This property is, in fact, independent on the scoring function s. We prove the exact distribution-free asymptotic law, only depending on score-generating function ϕ, similarly to the nonasymptotic guarantee of Prop. 72. In the article of Mann and Whitney (1947), the exact null distribution of their eponym statistic (proportional to the proposed statistic, with $\phi = Id$) was derived by means of a recurrence formulation. Later, Brus (1988); Chang (1992); Di Bucchianico (1999) for instance, proved this relation and provided a closed form for the moments of the statistic using combinatorial techniques. Nevertheless, for large values of N (greater than $n = m = 8$ in Mann and Whitney (1947)), the exact computation of the null distribution of linear rank statistics is very expensive (of factorial order). We propose estimating the null thanks to the asymptotic distribution for which its explicit parameters are detailed hereafter.

Proposition 75. *Suppose that Assumption 8 is fulfilled and $\phi : [0,1] \mapsto \mathbb{R}$ is nondecreasing. Then, under the null hypothesis $\mathscr{H}_0$ and for all $s \in \mathscr{S}_0$, the linear R-statistic $\widehat{W}^\phi_{n,m}(s)$ based on the two samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$, converges in distribution to:*

$$\frac{\widehat{W}^\phi_{n,m}(s) - \widehat{W}^0_\phi}{\sqrt{(N/n)\hat{\sigma}^2_\phi}} \xrightarrow[N\to\infty]{d} W \sim \mathscr{N}(0,1) \,, \tag{6.4.10}$$

where $\widehat{W}^0_\phi = \bar{\phi}_N = (1/N)\sum_{i\le N}\phi(i/(N+1))$ and $\hat{\sigma}^2_\phi = (1/N(N-1))\sum_{i\le N}(\phi(i/(N+1)) - \bar{\phi}_N)^2$.

Choosing $\phi = Id$, $n\widehat{W}^\phi_{n,m}(s)$ yields to $\widehat{W}^0_{Id} = n/2$ and $\hat{\sigma}^2_{Id} = nm/(12(N+1))$ and recovers the Wilcoxon (ranksum) statistic by noticing that is equals to $(N+1)\widehat{W}^{Id}_{n,m}$. This is used to tabulate the threshold values for the hypothesis test, depending on the size of samples n and m. Additionally, for very large values of N and if the estimator $\hat{\sigma}^2_\phi$ is consistent, one can use the asymptotic statistic to compute the threshold.

Proposition 76. *Suppose that Assumptions 8 and 9(ii) are fulfilled. Then, under the null hypothesis $\mathscr{H}_0$, for all $s \in \mathscr{S}_0$, the linear R-statistic $\widehat{W}_{n,m}^{\phi}(s)$ based on the two samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$, converges in distribution to:*

$$\sqrt{N}\left(\widehat{W}_{n,m}^{\phi}(s) - W_{\phi}^{0}\right) \xrightarrow[N\to\infty]{d} W \sim \mathscr{N}(0, \sigma_{\phi}^{2}) \,, \tag{6.4.11}$$

where $\sigma_{\phi}^{2} = (1/p)(\int_0^1 \phi^2 - (\int_0^1 \phi)^2)$.

6.5 Conclusion

This chapter applied *R*-processes, as introduced in Chapter 5, to the two-sample problem. By highlighting its adaptability, a two-stage procedure was proposed relying on bipartite ranking methods, leveraging on the interpretability of those statistics. Theoretical guarantees are proved, with nonasymptotic control of the deviation of the statistic under both statistical hypothesis, while the (non)asymptotic consistency of the overall procedure is guaranteed by results of Chapter 5. Importantly, this ranking approach to the two-sample problem allows for multiple practical adaptations to the practitioner. The numerical experiments highlighting the empirical properties of such method are detailed in Chapter 7.

6.6 Proofs

6.6.1 Proof of Formula (6.2.5)

Given two independent *r.v.* $X \sim G$ and $Y \sim H$, continuous, the proof is straightforward by noticing that

$$\int_{-\infty}^{\infty} \{H(t) - G(t)\} dH(t) = \frac{1}{2} - \int_{-\infty}^{\infty} G(t) dH(t)$$

so that

$$\mathrm{AUC}_{\mathrm{H,G}} = \mathbb{P}\{\mathrm{Y} \leq \mathrm{X}\} = \mathbb{E}[\mathbb{E}[\mathbb{I}\{\mathrm{Y} \leq \mathrm{X}\}] \mid \mathrm{Y}] = \mathbb{E}_{\mathrm{Y}\sim\mathrm{H}}[1 - \mathrm{G}(\mathrm{Y})] = \frac{1}{2} + \int_{-\infty}^{\infty} \{\mathrm{H}(\mathrm{t}) - \mathrm{G}(\mathrm{t})\} \mathrm{dH}(\mathrm{t}) \,.$$

6.6.2 Proof of Proposition 69

The equivalence between assertions (i) and (ii) can be straightforwardly deduced from the following result, proved in Clémençon and Vayatis (2009b) (see Corollary 5 and Proposition 6's proof therein), recalled here for the sake of clarity.

Lemma 77. *(Clémençon and Vayatis (2009b)) It holds with probability one:*

$$\Psi(\mathbf{Z}) = \frac{dG_{\Psi}}{dH_{\Psi}}(\Psi(\mathbf{Z})),$$

$\mathbf{Z}$ *denoting either* $\mathbf{X}$ *or* $\mathbf{Y}$.

The equivalence of assertion (ii) with the other assertions is immediate using formula (6.2.4).

6.6.3 Proof of Proposition 71

Let $s \in \mathscr{S}_0$ and suppose both assumptions 8 and 9(i) fulfilled. Following the footsteps of Chapter 5, the Taylor expansion of ϕ at order two at $u = N\widehat{F}_N(t)/(N+1)$ around $F(t)$ leads to an *a.s.* decomposition of the statistic $\widehat{W}^{\phi}_{n,m}$ when choosing $t = X_i$ and summing for all $i \leq n$, see Eq. (5.6.2) and (5.6.3) in Chapter 5. By writing the first order term in a sum of U-statistics, and applying next Hoeffding's decomposition yields to the following:

$$\widehat{W}^{\phi}_{n,m} - W_{\phi} = \widehat{W}_{\phi} - W_{\phi} + \frac{1}{n}\left(\widehat{V}^X_n - \mathbb{E}\left[\widehat{V}^X_n\right]\right) + \frac{1}{n}\left(\widehat{V}^Y_m - \mathbb{E}\left[\widehat{V}^Y_m\right]\right) + \frac{1}{n}\mathscr{R}_{n,m}\,, \tag{6.6.1}$$

where

$$\begin{aligned}
\widehat{W}_{\phi}(s) &= \frac{1}{n}\sum_{i=1}^{n}(\phi \circ F_s)\left(s(\mathbf{X}_i)\right)\,,\\
\widehat{V}^X_n(s) &= \frac{n-1}{N+1}\sum_{i=1}^{n}\int_{s(\mathbf{X}_i)}^{+\infty}(\phi' \circ F_s)(u)dG_s(u)\,,\\
\widehat{V}^Y_m(s) &= \frac{n}{N+1}\sum_{j=1}^{m}\int_{s(\mathbf{Y}_j)}^{+\infty}(\phi' \circ F_s)(u)dG_s(u)\,.
\end{aligned}$$

The remainder statistic $\mathscr{R}_{n,m}$ results from the integral Taylor-Lagrange term, and additional higher order statistics ($\mathscr{O}_{\mathbb{P}}(N^{-1})$) inherited from the Hoeffding decomposition. It is analyzed in Lemma 78 hereafter. First, we sequentially apply the tail inequality of Hoeffding (1963) to the empirical parts of the decomposition. Let $t > 0$,

$$\mathbb{P}\left\{\,|\widehat{W}_{\phi} - W_{\phi}| > t\right\} \leq 2\exp\left\{-\frac{2pNt^2}{\|\phi\|^2_{\infty}}\right\}\,, \tag{6.6.2}$$

$$\mathbb{P}\left\{\frac{1}{n}\,|\widehat{V}^X_n - \mathbb{E}[\widehat{V}^X_n]| > t\right\} \leq 2\exp\left\{-\frac{2Nt^2}{p\|\phi'\|^2_{\infty}}\right\}\,, \tag{6.6.3}$$

$$\mathbb{P}\left\{\frac{1}{n}\,|\widehat{V}^Y_m - \mathbb{E}[\widehat{V}^Y_m]| > t\right\} \leq 2\exp\left\{-\frac{2Nt^2}{(1-p)\|\phi'\|^2_{\infty}}\right\}\,. \tag{6.6.4}$$

For the remainder process, the following result is proved subsequently.

Lemma 78. *Suppose that the assumptions of Proposition 71 are satisfied. Then, for all $t > 0$ and $N \geq 2$, we have:*

$$\mathbb{P}\{|\mathscr{R}_{n,m}| > t\} \leq 12\begin{cases}\exp\left\{-\dfrac{Nt}{12\kappa_p\|\phi''\|_{\infty}}\right\}, & \text{if } Nt \geq 128\|\phi'\|^2_{\infty}/(p\|\phi''\|_{\infty})\\ \exp\left\{-\dfrac{\alpha_p N^2t^2}{512\|\phi'\|^2_{\infty}}\right\} & \text{otherwise}\end{cases}\,, \tag{6.6.5}$$

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

Then, the uniform bound yields the result where the statistic is valued at $\hat{s}$.

PROOF.(Lemma 78) Gathering the equations (B.4,8,15) of Clémençon et al. (2021), the remainder process is decomposed as

$$|\mathscr{R}_{n,m}| \leq |\widehat{R}_{n,m}| + p^2N|U_n| + p(1-p)N|U_{n,m}| + |\widehat{T}_{n,m}|\,, \tag{6.6.6}$$

We sequentially bound in probability each statistic, where $\widehat{R}_{n,m}$ encompasses bounded *i.i.d.* averages based on $\{X_1,\ldots,X_n\}$ as detailed below. $U_n(k)$ (*resp.* $U_{n,m}(\ell)$) is a one-sample degenerate U-statistic of order 2 based on $\{X_1,\ldots,X_n\}$ (*resp.* two-sample of degree $(1,1)$ based on the two samples $\{X_1,\ldots,X_n\},\{Y_1,\ldots,Y_m\}$). $\widehat{T}_{n,m}$ is the Taylor-Lagrange integral remainder of the expansion of ϕ at order 2. First, $\widehat{R}_{n,m}$ can be upperbounded as follows (see Eq. (6.6.7) in Chap. 5)

$$|\widehat{R}_{n,m}| \leq \frac{1}{N}\left|\sum_{i=1}^{n} G(X_i)\phi'\circ F(X_i) - \mathbb{E}[G(X_i)\phi'\circ F(X_i)]\right| + \frac{1}{N}\left|\sum_{i=1}^{n} H(X_i)\phi'\circ F(X_i) - \mathbb{E}[H(X_i)\phi'\circ F(X_i)]\right| . \quad (6.6.7)$$

By noticing that the variations can be bounded by $\|\phi'\|_\infty$, Hoeffding's inequality with the union bound directy yield, for $t>0$

$$\mathbb{P}\left\{(1/n)|\widehat{R}_{n,m}| > t/4\right\} \leq 4\exp\left\{-\frac{pN^3t^2}{32\|\phi'\|_\infty^2}\right\} . \quad (6.6.8)$$

For the two degenerate U-processes, Lemma 47 of Chap. 4 (Clémençon et al. (2021)) is applied, relying on Chernoff's and symmetrization methods.

$$\mathbb{P}\{p(1-p)(N/n)|U_{n,m}| > t/4\} \leq 2\exp\left\{-\frac{pN^2t^2}{512(1-p)\|\phi'\|_\infty^2}\right\} , \quad (6.6.9)$$

and similarly by Lemma 3 in Nolan and Pollard (1987)

$$\mathbb{P}\left\{p^2(N/n)|U_n| > t/4\right\} \leq 2\exp\left\{-\frac{N^2t^2}{512\|\phi'\|_\infty^2}\right\} . \quad (6.6.10)$$

Lastly, from Eq. (5.6.5) in Chap. 5

$$\begin{aligned}(1/n)|\widehat{T}_{n,m}| &\leq \|\phi''\|_\infty\left(\sup_{t\in\mathbb{R}}\left(\widehat{F}_N(t)-F(t)\right)^2+\frac{1}{(N+1)^2}\right)\\ &\leq 3p^2\|\phi''\|_\infty\sup_{t\in\mathbb{R}}\left(\widehat{G}_n(t)-G(t)\right)^2+3(1-p)^2\|\phi''\|_\infty\sup_{t\in\mathbb{R}}\left(\widehat{H}_m(t)-H(t)\right)^2+\frac{13\|\phi''\|_\infty}{N^2} \quad (6.6.11)\end{aligned}$$

By Dvoretzky–Kiefer–Wolfowitz inequality and noticing that the third term is rapidly negligible, we obtain

$$\mathbb{P}\left\{(1/n)|\widehat{T}_{n,m}| > t/4\right\} \leq 4\exp\left\{-\frac{Nt}{12\kappa_p\|\phi''\|_\infty}\right\} , \quad (6.6.12)$$

where $\kappa_p = \max(p,1-p)$. Applying the union bound to Eq. (6.6.8),(6.6.9),(6.6.10),(6.6.12) concludes the proof.

□

7 | Numerical Experiments

Abstract. This chapter gathers a series of numerical experiments, performed on synthetic data, for testing the procedures proposed in both Chapters 5 and 6. First, a deterministic algorithm based on a vanilla gradient ascent is detailed to maximize the smoothed version of the W_ϕ-performance criterion in the context of bipartite ranking. Additionally, we compare its performance through various choices of score-generating functions ϕ. Then, a panel of experiments are proposed for the high-dimensional two-sample problem with multiple choices of learning-to-rank state-of-the-art algorithms. The empirical type-I and type-II errors are compared for all algorithms to three classical test statistics: Maximum Mean Discrepancy (MMD, Gretton et al. (2012a)), Energy extension with the Euclidean norm (Energy, Szekely and Rizzo (2004)), Wald-Wolfowitz runs (FR, Friedman and Rafsky (1979)). All experiments are based on explicit probabilistic models that are detailed. We accompany these results with open access online codes available at `https://github.com/MyrtoLimnios`.

Contents

7.1 A deterministic approach to bipartite ranking

This section aims to empirically illustrate various points highlighted by the theoretical analysis carried out in Chapter 5 in the context of bipartite ranking. In particular, the capacity of ranking rules obtained by maximization of empirical W_ϕ-performance measures to generalize well and the impact of the choice of the score generating function ϕ on ranking performance from the perspective of ROC analysis. Some practical issues, concerning the maximization of smoothed versions of the empirical W_ϕ-performance criterion, are also discussed through numerical experiments. Additional experimental results can be found in the Appendix section 7.1.3. All experiments displayed in this section can be reproduced using the code available at `https://github.com/MyrtoLimnios/grad_2sample`.

7.1.1 A gradient-based algorithmic approach

We start by describing the gradient ascent method (GA) used in the experiments in order to maximize the smoothed criterion (5.4.10) obtained by kernel smoothing over the class of scoring functions $\mathscr{S}_0$ considered, as proposed in section 5.4.2, see Algorithm 3. Precisely, suppose that $\mathscr{S}_0$ is a parametric class, indexed by a parameter space $\Theta \subset \mathbb{R}^d$ with $d \geq 1$ say: $\mathscr{S}_0 = \{s_\theta : \mathscr{X} \to \mathbb{R},\ \theta \in \Theta\}$. Assume also that, for all $z \in \mathscr{X}$, the mapping $\theta \in \Theta \mapsto s_\theta(z)$ is of class $\mathscr{C}^1$ (*i.e.* continuously differentiable) with gradient $\partial_\theta s_\theta(z)$ and that the score-generating function ϕ fulfills Assumption 4. The gradient of the smoothed ranking performance measure of s_θ *w.r.t.* to the parameter θ, is given by: for all $\theta \in \Theta$, $h > 0$,

$$\nabla_\theta \left(\widehat{W}^{\phi}_{n,m,h}(s_\theta) \right) = \sum_{i=1}^{n} \phi' \left(\widehat{F}_{s_\theta,N,h}(s_\theta(\mathbf{X}_i)) \right) \nabla_\theta \left(\widehat{F}_{s_\theta,N,h}(s_\theta(\mathbf{X}_i)) \right) , \tag{7.1.1}$$

where the gradient of $\widehat{F}_{s_\theta,N,h}(s_\theta(z))$ *w.r.t.* to θ is:

$$\begin{aligned} \nabla_\theta \left(\widehat{F}_{s_\theta,N,h}(s_\theta(z)) \right) = \frac{1}{Nh} \sum_{i=1}^{n} K \left(\frac{s_\theta(z) - s_\theta(\mathbf{X}_i)}{h} \right) (\partial_\theta s_\theta(z) - \partial_\theta s_\theta(\mathbf{X}_i)) \\ + \frac{1}{Nh} \sum_{j=1}^{m} K \left(\frac{s_\theta(z) - s_\theta(\mathbf{Y}_j)}{h} \right) (\partial_\theta s_\theta(z) - \partial_\theta s_\theta(\mathbf{Y}_j)) , \end{aligned} \tag{7.1.2}$$

for any $z \in \mathscr{Z}$, using the fact that $\kappa' = K$.

Algorithm 3: Gradient Ascent for W-ranking performance maximization

Data: Independent *i.i.d.* samples $\{\mathbf{X}_i\}_{i\leq n}$ and $\{\mathbf{Y}_j\}_{j\leq m}$.
Input: Score-generating function ϕ, kernel K, bandwidth $h > 0$, number of iterations $T \geq 1$, step size $\eta > 0$.
Result: Scoring rule $s_{\widehat{\theta}_{n,m}}(z)$.

Choose the initial point $\theta^{(0)}$ in Θ;
for $t = 0, \ldots, T-1$ **do**
  compute the gradient estimate $\nabla_\theta \left(\widehat{W}^{\phi}_{n,m,h}(s_{\theta^{(t)}}) \right)$;
  update the parameter $\theta^{(t+1)} = \theta^{(t)} + \eta \nabla_\theta \left(\widehat{W}^{\phi}_{n,m,h}(s_{\theta^{(t)}}) \right)$;
end
Set $\widehat{\theta}_{n,m} = \theta^{(T)}$.

In practice, the iterations are continued until the order of magnitude of the variations $||\theta^{(t+1)} - \theta^{(t)}||$ becomes negligible. Then, the approximate maximum $s_{\widehat{\theta}_{n,m}}(z)$ output by Algorithm 3 is next used to rank test data. Averages over several Monte-Carlo replications are computed in order to produce the results displayed in Subsection 7.1.3.

7.1.2 Synthetic data generation

We now describe the data generating models used in the simulation experiments, as well as the parametric class of scoring functions, which the learning algorithm previously described is applied to.

Score-generating functions. To illustrate the importance of the function ϕ in the W_ϕ-performance ranking criterion, we successively consider $\phi_{MWW}(u) = u$ (MWW), $\phi_{Pol}(u) = u^q$, $q \in \mathbb{N}^*$ (Pol, Rudin (2006)) and $\phi_{RTB}(u) = \text{SoftPlus}(u - u_0) + u_0\text{Sigmoid}(u - u_0)$, $u_0 \in (0,1)$ (RTB, smoothed version of Clémençon and Vayatis (2007)), where the activation functions are defined by: $\text{SoftPlus}(u) = (1/\beta)\log(1+\exp(\beta u))$ and $\text{Sigmoid}(u) = 1/(1+\exp(-\lambda u))$, $\beta, \lambda > 0$ being hyperparameters to fit and control the derivative's slope.

Probabilistic models. Two classic two-sample statistical models are used here, namely the location and the scale models, where both samples are drawn from multivariate Gaussian distributions. We denote by $S_d^+(\mathbb{R})$ the set of positive definite matrices of dimension $d \times d$, by $\mathbb{I}_d$ the identity matrix.

Location model. Inspired by the optimality properties of linear rank statistics regarding shift detection in the univariate setup (*cf* Subsection 5.2.1), the model considered stipulates that $\mathbf{X} \sim \mathcal{N}_d(\mu_X, \Sigma)$ and $\mathbf{Y} \sim \mathcal{N}_d(\mu_Y, \Sigma)$ where $\Sigma \in S_d^+(\mathbb{R})$ and the mean/location parameters μ_X and μ_Y differ. The Algorithm 3 is implemented here with $\mathscr{Z} = \mathbb{R}^d = \Theta$ and $\mathscr{S}_0 = \{s_\theta(\cdot) = \langle \cdot, \theta \rangle, \ \ \theta \in \Theta\}$ as class of scoring functions, where $\langle \cdot, \cdot \rangle$ denotes the Euclidean scalar product on the feature space $\mathbb{R}^d$, and consequently exhibits no bias caused by the model. Indeed, by computing the loglikelihood ratio, one may easily check that the function $\langle \theta^*, \cdot \rangle$, where $\theta^* = \Sigma^{-1}(\mu_X - \mu_Y)$, is an optimal scoring function for the related bipartite ranking problem. Denoting by $\Delta(t) = (1/\sqrt{2\pi}) \int_{-\infty}^{t} \exp(-u^2/2)du$, $t \in \mathbb{R}$,

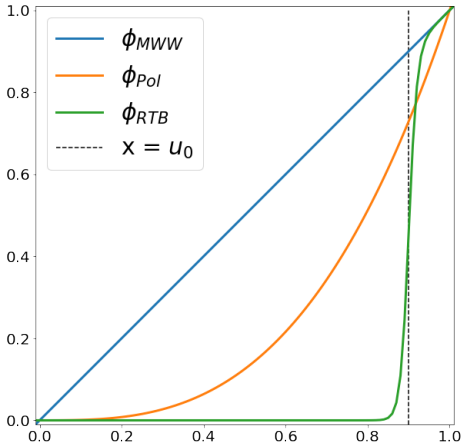


Figure 7.1. Curves of the three score-generating functions under study: $\phi_{MWW}(u) = u$ in blue, $\phi_{Pol}(u) = u^3$ in orange, $\phi_{RTB}(u) = \text{SoftPlus}(u - u_0) + u_0\text{Sigmoid}(u - u_0)$ the smoothed version of $u \mapsto u\mathbb{I}\{u \geq u_0\}$ in green, vertical line at $x = u_0$ in black.

the *c.d.f.* of the centered standard univariate Gaussian distribution, one may immediately check that the optimal ROC curve is given by:

$$\forall \alpha \in (0,1),\ \ \text{ROC}^*(\alpha) = 1 - \Delta\left(\Delta^{-1}(1-\alpha) + \sqrt{(\mu_X - \mu_Y)^T \Sigma^{-1}(\mu_X - \mu_Y)}\right).$$

Three levels of difficulty are tested through the implementations Loc1, Loc2 and Loc3. The nearly diagonal covariance matrix of the three models has its eigenvalues in $[0.5, 1.5]$ and $\mu_X = (1+\varepsilon)\mu_Y$ with $\varepsilon = 0.10$ (*resp.* $\varepsilon = 0.20$ and $\varepsilon = 0.30$) for Loc1 (*resp.* Loc2 and Loc3). The empirical ROC curves over the test pooled samples and additional curves are depicted in Fig. 7.5, 7.2, 7.5 for *resp.* Loc1, 2 and 3. The averaged test ROC curves and the *best* one among those produced through the replications made are gathered for the three models in Fig. 7.3. In Fig. 7.4, the evolution of the averaged empirical value of the W_ϕ-criteria on the train set during the algorithm is computed. Fig. 7.7 shows the results for Loc2 and 3 for three different parameters of the RTB model with $u_0 \in \{0.70,\ 0.90,\ 0.95\}$.

Scale model. Consider now the situation where $\mathbf{X} \sim \mathcal{N}_d(\mu, \Sigma_X)$ and $\mathbf{Y} \sim \mathcal{N}_d(\mu, \Sigma_Y)$, the distributions having the same location vector $\mu \in \mathbb{R}^d$ but different scale parameters Σ_X and Σ_Y in $S_d^+(\mathbb{R})$. The Algorithm 3 is implemented with $\mathscr{X} = \mathbb{R}^d$, $\Theta = S_d^+(\mathbb{R})$ and $\mathscr{S}_0 = \{s_\theta(z) = \langle z, \theta^{-1} z\rangle,\ \text{ for all } \ z \in \mathscr{X},\ \ \theta \in \Theta\}$, with the notations previously introduced. By computing the likelihood ratio, one immediately checks that $s_{\theta^*}(\cdot)$, with $\theta^* = \Sigma_X^{-1} - \Sigma_Y^{-1}$, is an optimal scoring function for the related scale model. For models Scale1, Scale2 and Scale3, observations are centered, $\Sigma_Y = \mathbb{I}_d$ and $\Sigma_X = \mathbb{I}_d + (\varepsilon/d)H$, where ε is taken equal to 0.70, 0.80 and 0.90 respectively and H a $d \times d$ symmetric matrix with real entries such that all the eigenvalues of $\Sigma_X \in S_d^+(\mathbb{R})$ are close to 1.

Similar to the location models, the empirical ROC curves over the test pooled samples and additional curves are depicted in Fig. 7.2, 7.6, 7.6 for *resp.* Scale1, 2 and 3. The averaged test ROC curves and the *best* one among those produced through the replications made are gathered for the three models in Fig. 7.3. In Fig. 7.4, the evolution of the averaged empirical value of the W_ϕ-criteria on the train set during the Algorithm is computed. Fig. 7.7 shows the results for Scale2 for three different parameters of the RTB model with $u_0 \in \{0.60,\ 0.70,\ 0.80\}$.

Experimental parameters. In all the experiments below, the pooled train sample is balanced ($p = 1/2$), *i.e.* $n = m = 150$ and the dimension of the feature space is $d = 15$. Similarly for the test sample with $n = m = 10^6$ and $d = 15$. Concerning the score-generating functions, we consider $q = 3$ (Pol) and $u_0 = 0.9$ (RTB). We use the Gaussian smoothing kernel $K(u) = (1/\sqrt{2\pi})\exp\{-u^2/2\}$ with a bandwidth $h = N^{-1/5}$, yielding an (asymptotically) optimal trade-off between bias and variance for the smoothed estimator of the *c.d.f.* involved in the criterion, see *e.g.* Girard and Saracco (2014). For completeness, the impact of the choice of the smoothing bandwidth h is investigated in Appendix 7.1.3: in short, a too large value for h flattens the criterion and significantly slows down the convergence, while too small values make the gradient ascent algorithm very unstable. Algorithm 3 is implemented with $T = 50$ and a learning step size η of order $1/\sqrt{T}$. For each model, $B = 50$ Monte-Carlo replications of the train pooled sample are generated. Based on the latter, the learning algorithm is implemented B times and an average curve and a standard deviation based on the test ROC curves thus obtained are computed for each model in a pointwise manner.

Evaluation of the criteria. In order to evaluate the performance of a scoring function produced by an early-stopped version of Algorithm 3 depending on the score-generating function chosen, it is used to score the observations of a large test sample. Using a Monte Carlo procedure, this is replicated for B independent training datasets and the corresponding B test ROC curves are computed and are compared to that of the optimal scoring function $s_{\theta^*}(z)$. In what follows, 'average/best/worst' criterion values (respectively, test ROC curves, scoring functions) refer to the values of the criterion taken over the B Monte Carlo replications. Particular attention is paid to the behavior of the test ROC curves near the origin, which reflects the ranking performance for the instances with highest score values.

7.1.3 Results and discussion

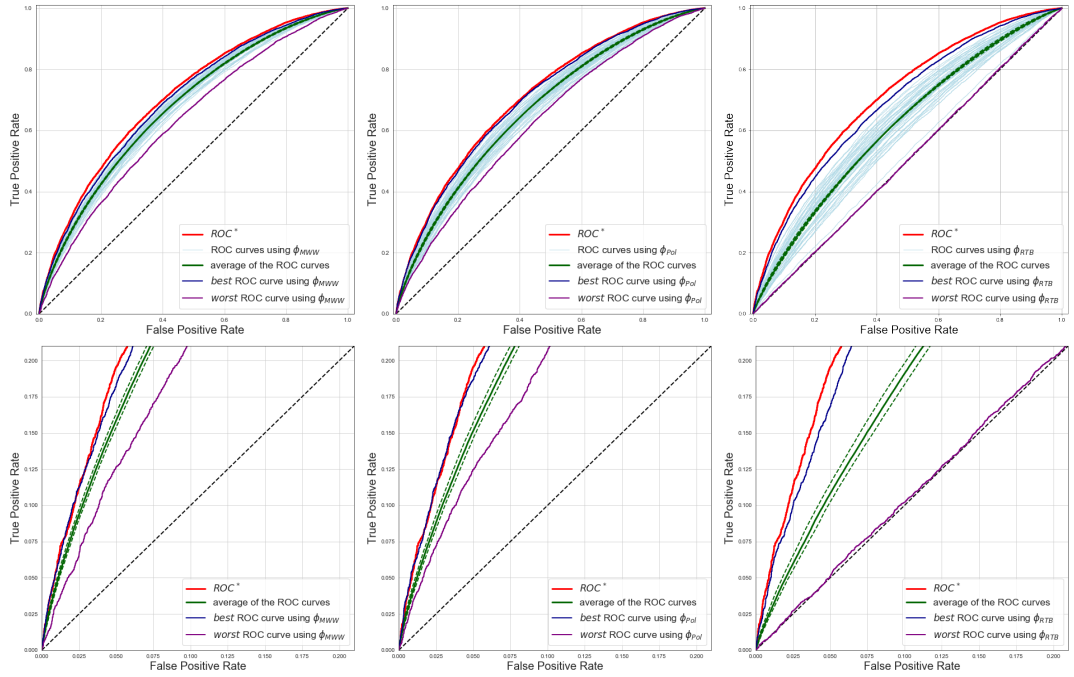
We now analyze the experimental results, by commenting on the test ROC curves obtained after learning the scoring functions, using the early-stopped version of the Algorithm 3 described above, that maximize the chosen (smoothed variant of the) W_ϕ-performance measure: MWW, Pol and RTB. We compare them with ROC*. All the experiments were run using Python.

For both the location and scale models, we ran the algorithm for three increasing levels of difficulty defined by the decreasing value of the parameter ε. Figures 7.3 (location and scale) show that the three methods (MWW, Pol, RTB) learn an empirical parameter $\widehat{\theta}_{n,m}$ such that the corresponding ROC curve gets close to ROC* (red curves) and the more ε increases and the more the scoring rule learned generalizes well. Fig. 7.4 (location and scale) reveal the monotonicity of the evolution of the empirical criteria, as the number of iterative steps of Algorithm 3 increases. Unsurprisingly, all the results show an increasing ability to learn a scoring function that maximizes the three W_ϕ-performance measures, as ε increases (*i.e.* when the distribution G and H are significantly more different from each other).

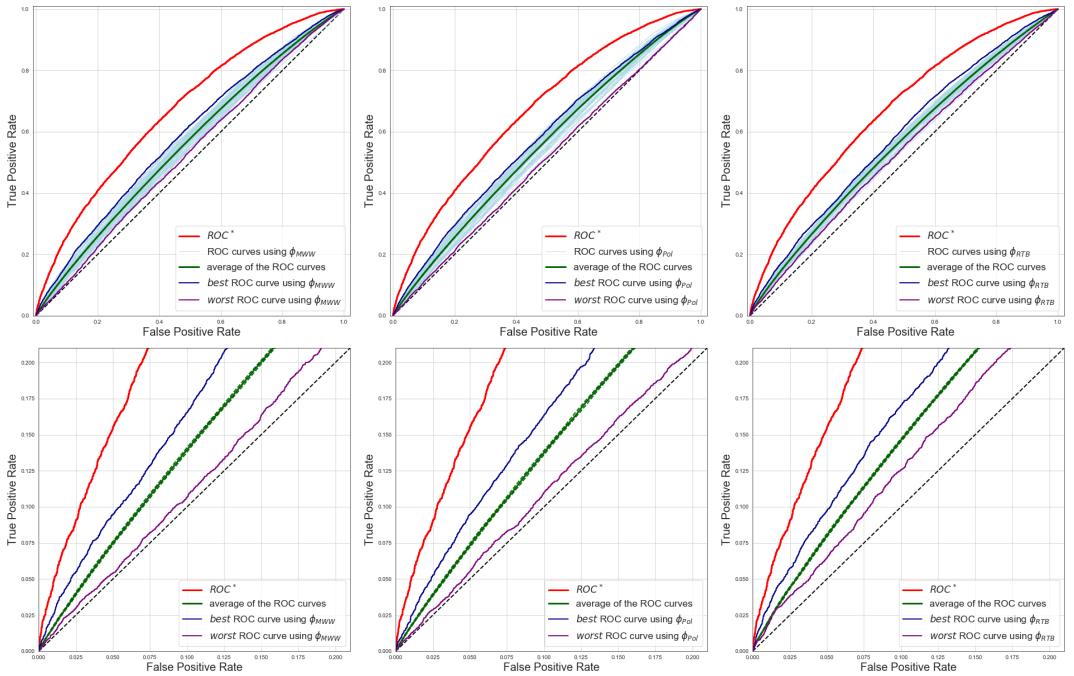
Analyzing the average of the empirical ROC curves obtained, MWW performs better for the location model as its corresponding curve converges faster to ROC* for all ε. This phenomenon was expected due to the well-known high power of the related Mann-Whitney-Wilcoxon test statistic in this modeling. The aggregated ROC curve for the Pol method also performs well, while RTB's presents a low performance compared to MWW, see Fig. 7.3. Indeed, considering only the best ranked observations at each iteration in the learning procedure, does not always achieve a good scoring parameter and is enhanced by the early-stopped rule. It results in a higher variance and a larger

spectrum of the empirical curves both at the same time, see the light blue curves in Fig. 7.2.3. and 7.5.3. (Loc2 and Loc3). The slow convergence for the RTB method is illustrated with Loc1, where almost both samples are blended/coincide, for which only the ROC curves above the diagonal were kept. For the scale model, the aggregated ROC curves are comparable for the three methods with a slightly higher performance obtained by RTB and we note the faster convergence of the algorithm for this model, see Fig. 7.4.

Looking at the *best* ROC curves (dark blue lines), defined as those obtained by the scoring function minimizing the generalization error for each criterion, RTB yields to a scoring function that generalizes best for most of the models. In particular, when focussing on the 'best' instances in the learning procedure, the obtained empirical scoring functions have higher performance at the beginning of the ROC curves, see the zoomed plots. Also, choosing the optimal proportion $1-u_0$ of observations to consider for the score-generating function results in different performance measures. Figure 7.7 gathers the resulting plots for models Loc2 and 3 with u_0 in $\{0.7, 0.9, 0.95\}$ while Fig. 7.7 depicts the scale model 2 with u_0 in $\{0.6, 0.7, 0.8\}$ and a higher number of loops $T = 70$. Considering the *best* ROC curves for all models shows that when u_0 tends to one, the beginning of the curve is accurately learned. Incidentally, note that the proportion of observations considered has to be large enough, so that the optimization algorithm performs well.

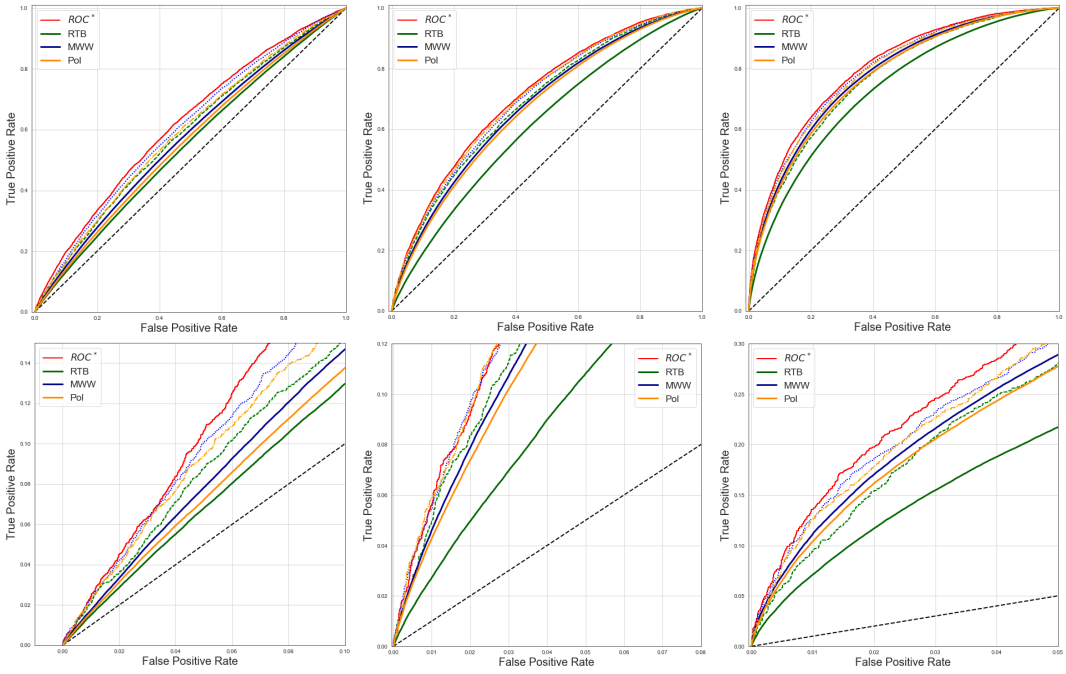


1. Loc2, $\phi_{MWW}(u) = u$ 2. Loc2, $\phi_{Pol}(u) = u^3$ 3. Loc2, $\phi_{RTB}(u) = u\mathbb{I}\{u \geq 0.9\}$

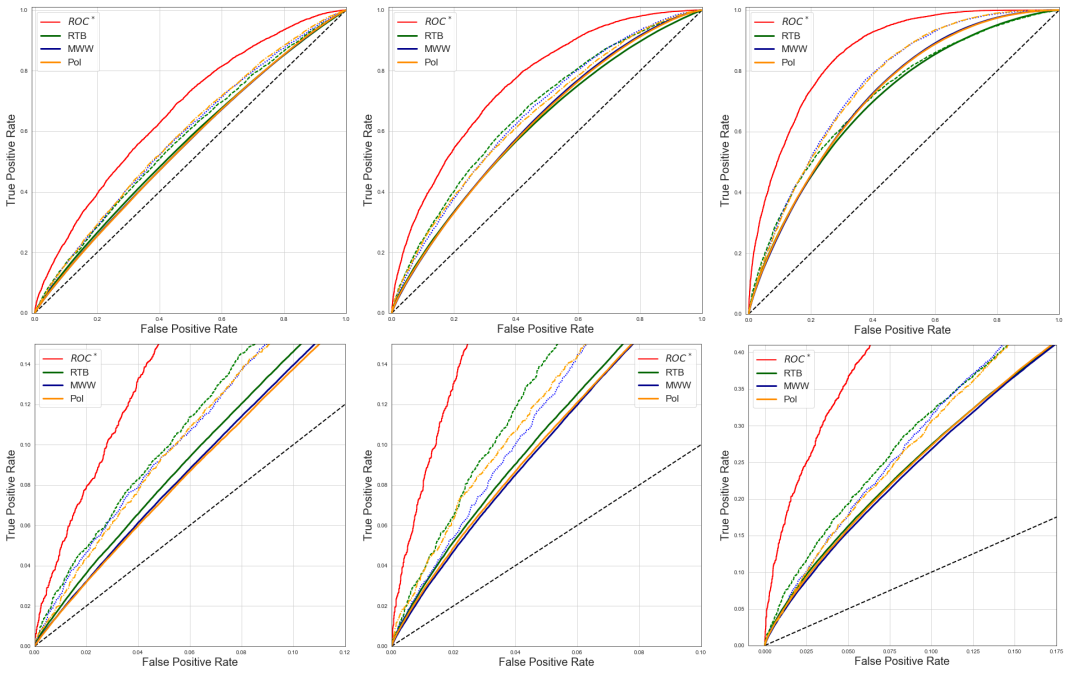


4. Scale1, $\phi_{MWW}(u) = u$ 5. Scale1, $\phi_{Pol}(u) = u^3$ 6. Scale1, $\phi_{RTB}(u) = u\mathbb{I}\{u \geq 0.9\}$

Figure 7.2. Empirical ROC curves and average ROC curve for Loc2 (1-3) ($\varepsilon = 0.20$) and for Scale1 (4-6) ($\varepsilon = 0.70$). Samples are drawn from multivariate Gaussian distributions according to section 7.1.2,scored with early-stopped GA algorithm's optimal parameter for the class of scoring functions. Hyperparameters: $u_0 = 0.9$, $q = 3$, $B = 50$, $T = 50$. Parameters for the training set: $n = m = 150$; $d = 15$; for the testing set: $n = m = 10^6$; $d = 15$. Figures $1,2,3$ correspond *resp.* to the models MMW, Pol, RTB. Light blue curves are the $B(= 50)$ ROC curves that are averaged in green (solid line) with $+/-$ its standard deviation (dashed green lines). The dark blue and purple curves correspond to the best and worst scoring functions in the sense of minimization and maximization of the generalization error among the B curves. The red curve corresponds to ROC*.

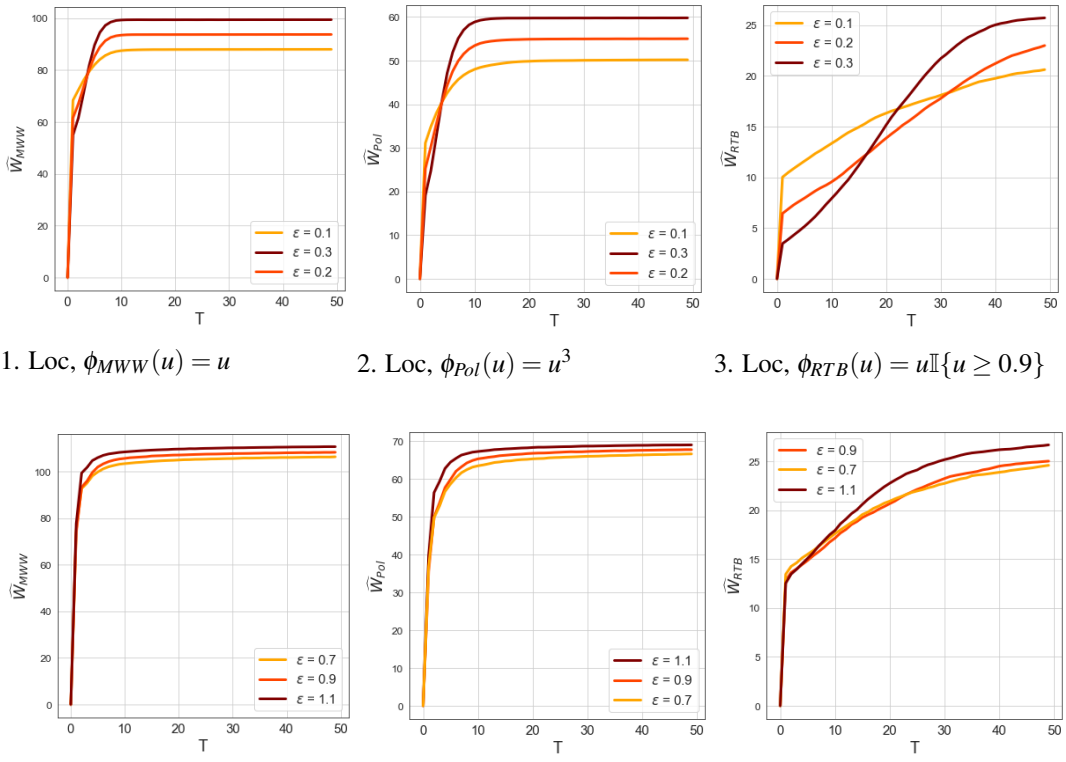


1. Loc1, $\varepsilon = 0.10$ 2. Loc2, $\varepsilon = 0.20$ 3. Loc3, $\varepsilon = 0.30$



4. Scale1, $\varepsilon = 0.70$ 5. Scale2, $\varepsilon = 0.90$ 6. Scale3, $\varepsilon = 1.10$

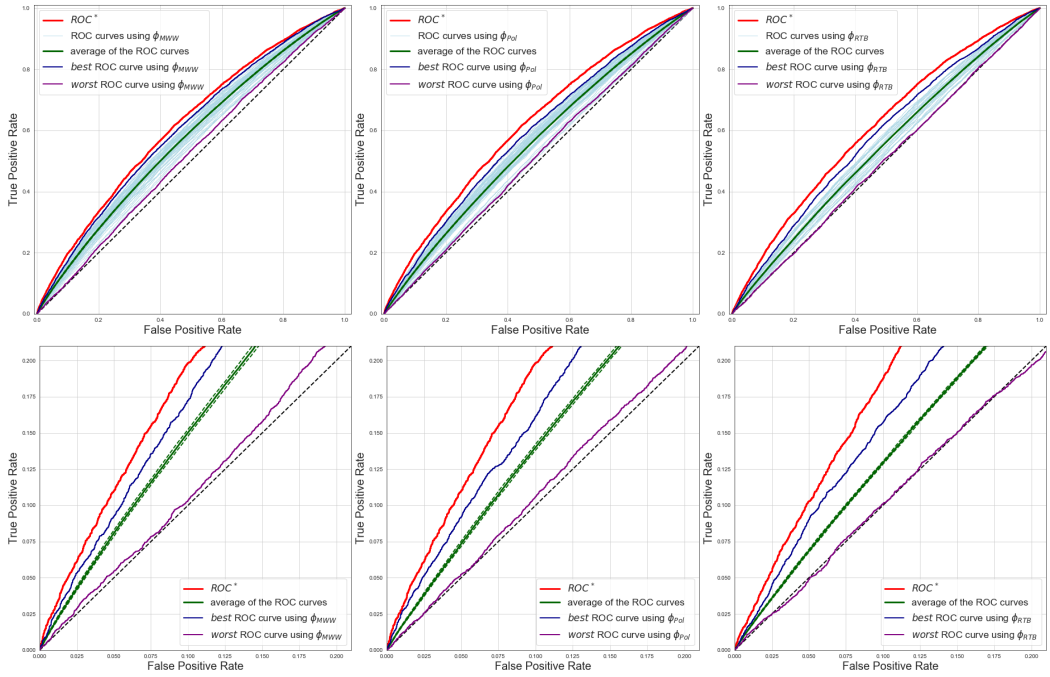
Figure 7.3. Average of the ROC curves (solid line), *best* ROC curves (dashed line) for the three location models Loc1 (1.), Loc2 (2.) and Loc3 (3.) and for the scale models Scale1 (4.), Scale2 (5.) and Scale3 (6.). In blue for MWW, orange for Pol, green for RTB, red for ROC*. Samples are drawn from multivariate Gaussian distributions according to section 7.1.2, scored with early-stopped GA algorithm's optimal parameter for the class of scoring functions and averaged after $B = 50$ loops. Hyperparameters: $u_0 = 0.9$; $q = 3$, $B = 50$, $T = 50$. Parameters for the training set: $n = m = 150$; $d = 15$; for the testing set: $n = m = 10^6$; $d = 15$.



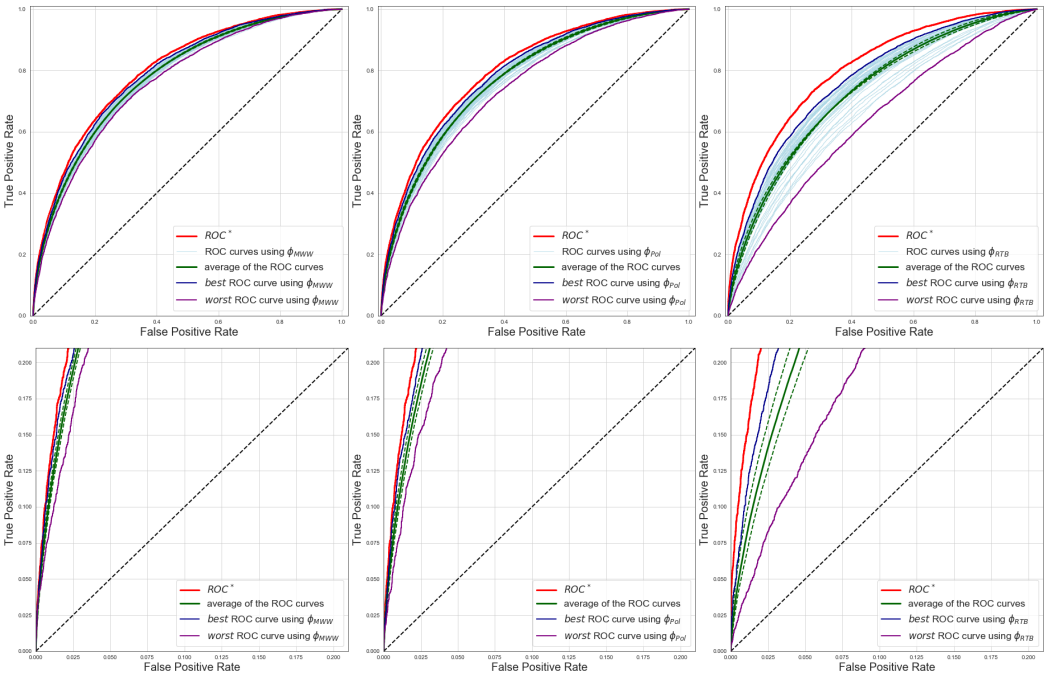
1. Loc, $\phi_{MWW}(u) = u$ 2. Loc, $\phi_{Pol}(u) = u^3$ 3. Loc, $\phi_{RTB}(u) = u\mathbb{I}\{u \geq 0.9\}$

4. Scale, $\phi_{MWW}(u) = u$ 5. Scale, $\phi_{Pol}(u) = u^3$ 6. Scale, $\phi_{RTB}(u) = u\mathbb{I}\{u \geq 0.9\}$

Figure 7.4. Average of the empirical W_ϕ-ranking performance measure over the $B = 50$ loops for the three location models Loc1, Loc2 and Loc3 and for three scale models Scale1 , Scale2 and Scale3. Samples are drawn from multivariate Gaussian distributions according to section 7.1.2, scored with early-stopped GA algorithm's optimal parameter for the class of scoring functions and averaged after $B = 50$ loops. Hyperparameters: $u_0 = 0.9$; $q = 3$, $B = 50$, $T = 50$. Parameters for the training set: $n = m = 150$; $d = 15$; for the testing set: $n = m = 10^6$; $d = 15$.

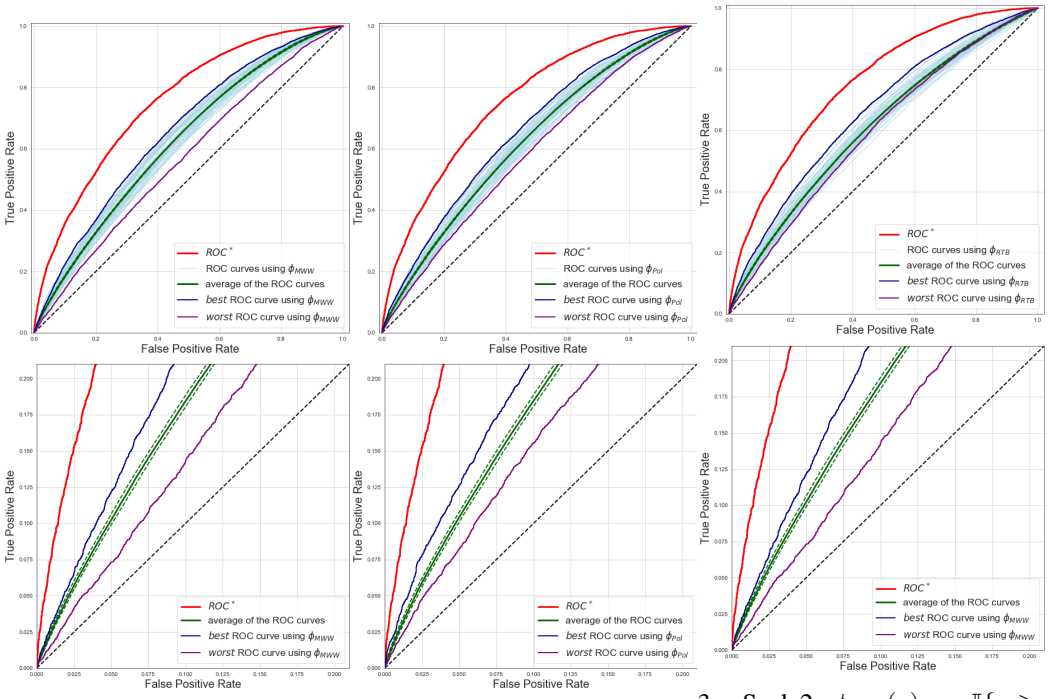


1. Loc1, $\phi_{MWW}(u) = u$ 2. Loc1, $\phi_{Pol}(u) = u^3$ 3. Loc1, $\phi_{RTB}(u) = u\mathbb{I}\{u \geq 0.9\}$

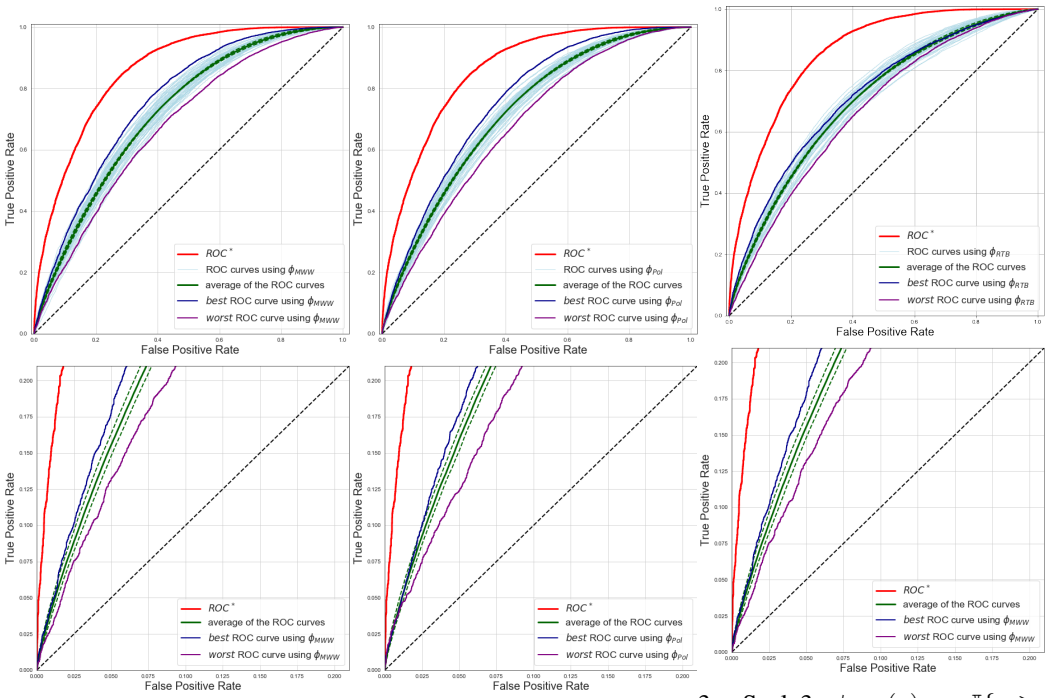


1. Loc3, $\phi_{MWW}(u) = u$ 2. Loc3, $\phi_{Pol}(u) = u^3$ 3. Loc3, $\phi_{RTB}(u) = u\mathbb{I}\{u \geq 0.9\}$

Figure 7.5. Empirical ROC curves and average ROC curve for Loc1 ($\varepsilon = 0.10$), Loc3 ($\varepsilon = 0.30$). Samples are drawn from multivariate Gaussian distributions according to section 7.1.2, scored with early-stopped GA algorithm's optimal parameter for the class of scoring functions. Hyperparameters: $u_0 = 0.9$, $q = 3$, $B = 50$, $T = 50$. Parameters for the training set: $n = m = 150$; $d = 15$; for the testing set: $n = m = 10^6$; $d = 15$. Figures 1,2,3 correspond *resp.* to the models MMW, Pol, RTB. Light blue curves are the $B(= 50)$ ROC curves that are averaged in green (solid line) with $+/-$ its standard deviation (dashed green lines). The dark blue and purple curves correspond to the best and worst scoring functions over the B Monte Carlo replications. The red curve corresponds to ROC*.



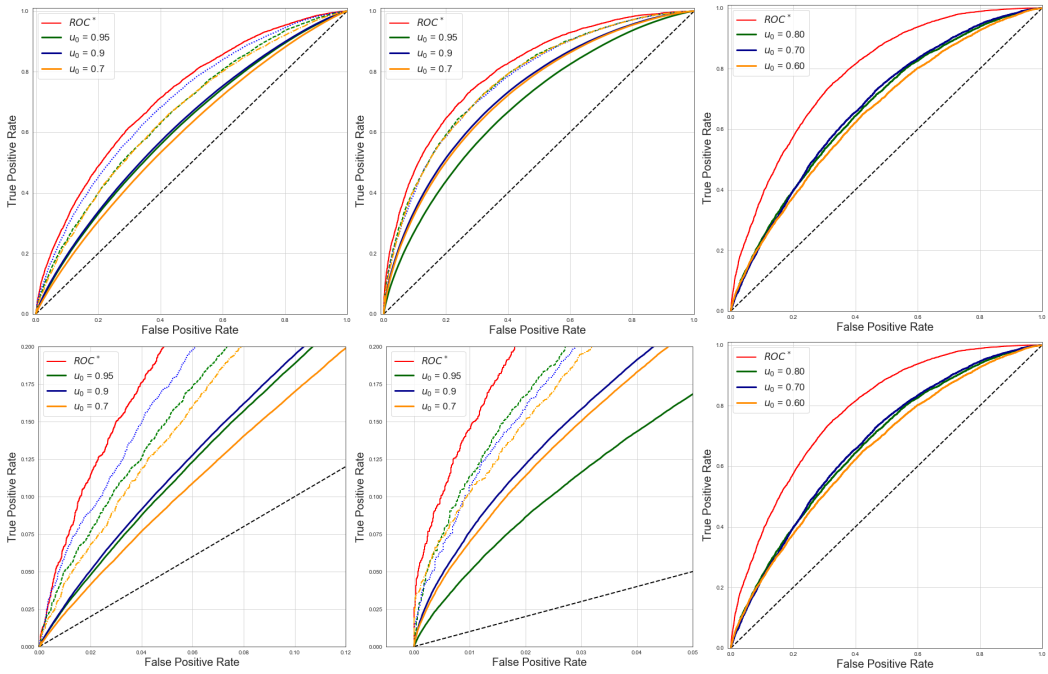
1. Scale2, $\phi_{MWW}(u) = u$ 2. Scale2, $\phi_{Pol}(u) = u^3$ 3. Scale2, $\phi_{RTB}(u) = u\mathbb{I}\{u \geq 0.9\}$



1. Scale3, $\phi_{MWW}(u) = u$ 2. Scale3, $\phi_{Pol}(u) = u^3$ 3. Scale3, $\phi_{RTB}(u) = u\mathbb{I}\{u \geq 0.9\}$

Figure 7.6. Empirical ROC curves and average ROC curve for Scale2 ($\varepsilon = 0.90$) and Scale3 ($\varepsilon = 1.10$). Samples are drawn from multivariate Gaussian distributions according to section 7.1.2, scored with early-stopped GA algorithm's optimal parameter for the class of scoring functions. Hyperparameters: $u_0 = 0.9$, $q = 3$, $B = 50$, $T = 50$. Parameters for the training set: $n = m = 150$; $d = 15$; for the testing set: $n = m = 10^6$; $d = 15$. Figures $1,2,3$ correspond *resp.* to the models MMW, Pol, RTB. Light blue curves are the $B(= 50)$ ROC curves that are averaged in green (solid line) with $+/-$ its standard deviation (dashed green lines). The dark blue and purple curves correspond to the best and worst scoring functions over the B Monte Carlo replications. The red curve corresponds to ROC*.

Additional results. This paragraph gathers some empirical results regarding the difference in performance of the W-criteria for the RTB score-generating function, when we vary the rate u_0, for both the location and the scale model in Fig. 7.7. Lastly, in order to highlight the effect of a too small/large value for the smoothing parameter h, Fig. 7.8 depicts the empirical results related to the location model (Loc1) when considering smoothed versions of the MWW criterion in the learning stage and varying h: the value of the test MWW criterion evaluated at the scoring function output by Algorithm 3 is plotted for each of the B Monte-Carlo replications (when the algorithm diverges, the value is set to zero by convention). As expected, it shows that the performance attained after a fixed number of iterations deteriorates in average for too large values of the bandwidth h (the criterion used in the learning stage flattens itself as h increases), whereas a greater unstability is observed when h is too small.



1. Loc2, $\varepsilon = 0.2$, $T = 50$ 2. Loc3, $\varepsilon = 0.3$, $T = 50$ 2. Scale2, $\varepsilon = 0.9$, $T = 70$

Figure 7.7. Comparison of three RTB models. Average of the ROC curves (solid line), *best* ROC curves (dashed line) for the two location models Loc2 and Loc3. Samples are drawn from multivariate Gaussian distributions according to section 7.1.2, scored with early-stopped GA algorithm's optimal parameter for the class of scoring functions and averaged after B loops. Hyperparameters: $B = 50$, $T = 50$ for the location models and $T = 70$ for Scale2. Parameters for the training set: $n = m = 150$; $d = 15$; for the testing set: $n = m = 10^6$; $d = 15$.

7.2 Ranking-based two-sample testing

In the context of the two-sample problem, this section provides a series of numerical experiments to discuss technical aspects involved in the proposed ranking-based procedure on simulated datasets. First, let two independent *i.i.d.* random samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$, with $n, m \in \mathbb{N}^*$, drawn from unknown G and H and valued in the measurable space $\mathscr{X}$. The goal is to test the hypothesis below for a fixed level $\alpha \in (0,1)$ and based on the two samples:

$$\mathscr{H}_0 : G = H \text{ against the alternative } \mathscr{H}_1 : G \neq H \,. \tag{7.2.1}$$

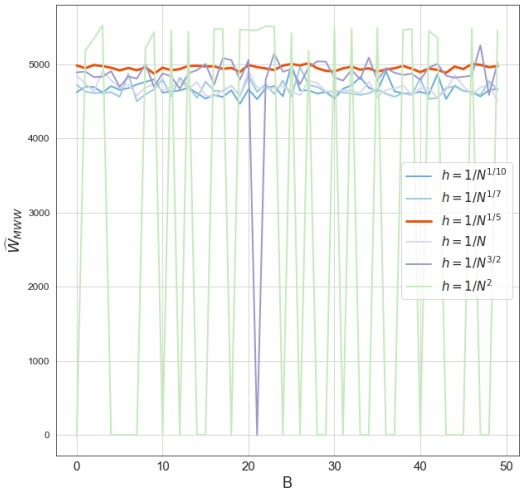


Figure 7.8. Test values of the W_ϕ-criterion for the experiment Loc1, with $\phi_{MWW}(u) = u$, for each Monte-Carlo replication, depending on the smoothing bandwidth used in the learning stage: $h \in \{0.1, 1/7, 0.2, 1, 1.5, 2\}$. The red curve corresponds to the bandwidth value proposed in Section 7.1.

In particular, the present objective is twofold: (i) to compare the performance of the proposed procedure to classic multivariate two-sample tests, for various bipartite-learning algorithms, (ii) to analyze its performance depending on the choice of the score-generating function ϕ. We first detail the algorithmic elements of the bipartite ranking step, defined to learn the optimal scoring function on the training dataset. Then the probabilistic models are presented, and the empirical results are discussed. All experiments displayed in this section can be reproduced using the code available at `https://github.com/MyrtoLimnios`.

7.2.1 Algorithms

We consider the two-stage testing procedure as summarized in 6.2 and briefly recall the framework. The initial two-sample dataset is composed of independent and *i.i.d.* samples $\mathscr{D}_{n,m} = \{\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}, \{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}\}$ of sizes n, $m \geq 2$ and valued in $\mathscr{X} \subset \mathbb{R}^d$, with $d \geq 2$. The technicalities of each step are sequentially detailed, followed by the benchmark comparison tests and finally the experimental parameters.

Bipartite ranking algorithms (*Step 1*). Consider an algorithm $\mathscr{A}$ trained on a subset $\mathscr{D}_{n',m'} \subset \mathscr{D}_{n,m}$, with $n' < n$, $m' < m$. For a given class $\mathscr{S}_0$, the goal is to learn the scoring function $\hat{s} \in \mathscr{S}_0$ that induces the optimal ranks of the random observations in the sense of the bipartite ranking loss criterion. As exposed in the Section 6.2.3, the objective loss usually boils down to maximizing the (tailored) empirical AUC. From the algorithmic perspective, this statistic is challenging due to *e.g.* the comparison of pairs, leading to quadratic complexity. However, many algorithms have been introduced to approximate the exact empirical loss, in particular in the context of pairwise learning-to-rank frameworks. They are designed to bridge the gap to pairwise models by generating all possible couples from the two samples and recovering the exact objective under the assumption of linear separation between the samples. We propose a modified version to fit best to the proposed statistic that generates only a particular class of pairs of instances. In particular, the linear `RankSVM` (see Joachims (2002)) with L_1 and L_2 losses (*resp.* rSVM1, rSVM2), as well as both `RankNN` (RNN, see Burges et al. (2005)) and `LambdaRankNN` (lRNN, see Burges et al. (2007)), are implemented. We also provide the algorithms for boosting methods, `RankBoost` and `AdaBoost`. The modified

algorithms are detailed in the Appendix section 7.4. These algorithms are particularly well fitted for large samples and in the nonparametric setting *i.e.* when the type of *transformation* between the two underlying probability laws is unknown. We also test classic algorithms for binary classification, namely the `LinearSVR` (SVR) and the logistic regression (LR) for the sake of algorithmic simplicity, available directly on the `sklearn` librairy of `Python`.

Remark 10. (ON THE EXACT MAXIMIZATION OF W_ϕ) *When few observations are available, such algorithms might not converge. We propose an exact bipartite ranking algorithm based on a gradient ascent optimization method. This relies on maximizing an empirical version of the smoothed W_ϕ-criterion, as introduced in Section 7.1. Although it requires parametrizing $\mathscr{S}_0$, we showed it recovers at best the* ROC *curve with different characteristics depending on the choice of ϕ. This is investigated in Appendix sections A.2 and A.3.*

Univariate two-sample rank tests (*Step 2*). Consider the optimal output $\hat{s}$ of the first step, such that the test subset $\mathscr{D}_{n'',m''}$ is scored and valued in $\mathbb{R}$. We implement the statistic $\widehat{W}^{\phi}_{n,m}(s)$, for $\phi_{MWW}(u) = u$ (MWW, Wilcoxon (1945)), $\phi_{RTB}(u) = u\mathbb{I}\{u \geq u_0\}$ for $u_0 \in (0,1)$ (RTB, Clémençon and Vayatis (2007)). For these choices of ϕ and for a given a level α, we evaluate our method *w.r.t.* the type-I error and the power of the obtained test statistics. Notice that the directional error (or type-III) is null when the underlying probabilistic model is known. For simulated datasets, we first consider both probabilities when the *dissimilarity/discrepancy* parameter $\varepsilon > 0$ varies with fixed design, then we fix the dissimilarity parameter but let the dimension d increase. Finally, for some experiments, the empirical ROC curves based on the scored test sample are plotted.

Null distribution approximation. The linear rank statistic being exactly distribution-free, its univariate distribution under the null hypothesis can be computed through a simple procedure, *c.f.* Section 6.4.3. Although this property, its algorithmic complexity is of factorial order thus requiring high computational capacity. Therefore, two distribution-free rules are implemented depending on the sample sizes n, m, following the explicit formulations derived in Sections 6.4.1 and 6.4.3. Of course, when the score-generating function ϕ corresponds to a classic univariate two-sample test, the `SciPy` open-access library available in `Python` can be used directly on the scored samples to perform the test.

Benchmark tests. We compare our results to three classic multivariate and nonparametric two-sample tests from the literature. The unbiased (quadratic) Maximum Mean Discrepancy (MMD) test with Gaussian kernels from Gretton et al. (2007, 2012a), the graph-based Wald-Wolfowitz runs test (FR) generalized to the multivariate setting in Friedman and Rafsky (1979), the metric-based Energy test (Energy), see Székely and Rizzo (2013). (Notice that both MMD and ED are not exactly distribution-free tests.) For the selected tests, we reviewed and updated open-access `Python` libraries to unify the testing framework, in particular to ensure that the estimation of the null distribution is common. We point out that, even if experimental models are parametric, we do not implement multivariate parametric tests, as those listed in Section 6.2.2.

Experimental parameters. For all the experiments, we consider the proportion for the train/test for each sample equals to $n'/n = 4/5$ and $n''/n = 1/5$ and $n, m = 500, 500$, the power is estimated *via* Monte-Carlo method and averaged over $B = 50$ replications, the significance level of the tests is set to $\alpha = 0.05$. The parameter of the RTB functions are $u_0 \in \{0.6, 0.7, 0.8, 0.9\}$. For the samples parameters, $n = m$, with $n = 500$, and $d > 1$ depending on the probabilistic model. For the benchmark tests, the null distribution is estimated with the permutation method set to $B_{perm} = 100$ and the

hyperparameter $\beta \in \{1e-3, 1e-2, 1e-1, 1, 5, 10, 15, 20, 25, 30, 1e2, 1e3\}$. The plots include the confidence interval for each point at level 95%. For a given experiment, the size of each sample is kept fixed at each train/test split. Also, the hyperparameters are not optimized.

Evaluation criteria. In order to evaluate the performance of the procedure for various ranking algorithms, we estimate the empirical type-I error, corresponding to $\varepsilon = 0$, and the empirical power, for multiple choices of $\varepsilon > 0$. Precisely, a Monte-Carlo procedure is performed, as detailed in Chapter 6, section 6.4.2, with $B \in \mathbb{N}^*$ replications as follows. For a given sample, the ranking algorithm of *Step 1* outputs the optimal scoring functions $\hat{s}$ on the training sample. Then for *Step 2*, the scores of the remaining two samples are computed and the homogeneity test is performed at risk α, *i.e.* $\Phi_\alpha^\phi(\mathscr{D}_{n'',m''}(\hat{s}))$, see Eq. (6.3.10). As a global evaluation criterion, the optimal test is selected using the minimax separation rate ρ, as detailed in Chapter 6, section 6.4.2. It aims to output the test achieving the higher power for the smaller ε in the sense of the Oracle test (AUC*, with the notation of the previous section 7.1), while controlling minimal type-I error.

7.2.2 Synthetic data generation

We illustrate the proposed class of rank-based test statistics performances through the following frameworks. Various location and scale models are implemented to compare the empirical results to the explicit oracle statistic. The additional parameters are detailed in the appendix section for clarity.

Location two-sample tests for Gaussian samples. The two samples $\mathbf{X} \sim \mathscr{N}_d(\mu_X, \Sigma)$ and $\mathbf{Y} \sim \mathscr{N}_d(\mu_Y, \Sigma)$ are drawn independently, with $\Sigma \in S_d^+(\mathbb{R})$, $\varepsilon \in \{0.0, 0.2, 0.3, 0.6\}$, as follows:

(L1) $\mu_Y = 0_d$, $\mu_X = (\varepsilon/\sqrt{d}) \times \mathbf{1}_d$. Two modelings for the covariance matrix where the first marginal is negatively correlated with all the others and for $2 \leq k \leq d$ the coordinates are : (L1$-$) mutually independent; (L1$+$) positively correlated.

The explicit covariance matrices Σ are displayed in Section 7.4. This class of models has an explicit solution of scoring class $\mathscr{S}^*$ corresponding of the linear functions satisfying $s_\theta(\cdot) = \langle \theta, \cdot \rangle$, $\theta \in \Theta \subset \mathbb{R}^d$ and of optimal parameter proportional to $\theta^* = \Sigma^{-1}(\mu_X - \mu_Y)$, see Clémençon et al. (2021). The true ROC curves (ROC*) are plotted in the Appendix section 7.4.

Scale two-sample tests for Gaussian samples. The two samples $\mathbf{X} \sim \mathscr{N}_d(0_d, \Sigma_X)$, and $\mathbf{Y} \sim \mathscr{N}_d(0_d, \Sigma_Y)$ are drawn independently with Σ_X, $\Sigma_Y \in S_d^+(\mathbb{R})$ as follows:

(S1) $\mathbf{X} \sim \mathscr{N}_d(0_d, \Sigma_X)$, and $\mathbf{Y} \sim \mathscr{N}_d(0_d, \mathbb{I}_d)$, with $\Sigma_X = \mathbb{I}_d + (\varepsilon/d)H$, H a symmetric invertible matrix of eigenvalues in $[0.5, 1.5]$, with $d \in \{50, 100\}$.

(S2) *Decreasing correlation.* $\mathbf{X} \sim \mathscr{N}_d(0_d, \Sigma_X)$, and $\mathbf{Y} \sim \mathscr{N}_d(0_d, \Sigma_Y)$, with $\Sigma_{X,i,j} = \alpha^{|i-j|}$, $\Sigma_{Y,i,j} = \beta^{|i-j|}$, for $i, j \leq d$, with $d \in \{3, 20\}$, $\beta = 0.2$ and $\alpha = \beta + \varepsilon$.

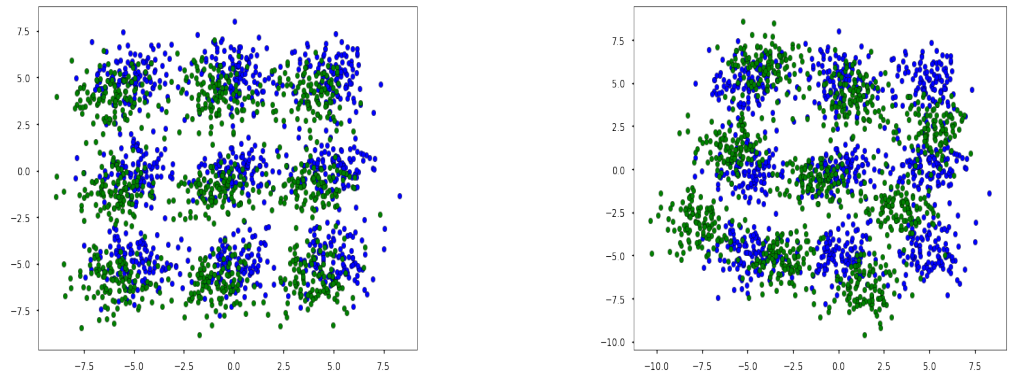
(S3) *Equi-correlated samples.* $\mathbf{X} \sim \mathscr{N}_d(0_d, \Sigma_X)$, and $\mathbf{Y} \sim \mathscr{N}_d(0_d, \Sigma_Y)$, with $\Sigma_X = (1-\alpha)\mathbb{I}_d + \alpha \mathbf{1}_d \mathbf{1}_d^T$, $\Sigma_Y = (1-\beta)\mathbb{I}_d + \beta \mathbf{1}_d \mathbf{1}_d^T$, with $d \in \{3, 20\}$, $\beta = 0.3$ and $\alpha = \beta + \varepsilon$.

The optimal class of scoring class $\mathscr{S}^*$ is defined by the set of quadratic functions satisfying $s_\theta(\cdot) = \langle \cdot, \theta^{-1} \cdot \rangle$, $\theta \in \Theta = S_d^+(\mathbb{R})$ and of optimal parameter proportional to $\theta^* = \Sigma_X^{-1} - \Sigma_Y^{-1}$, see Clémençon et al. (2021). The ROC* curves are plotted in the Appendix section 7.4.

Location and scale mixture models. We illustrate the case for both location and scale/rotation mixture models with samples drawn from Gaussian distributions. We use the blobs dataset as a classic example provided by the library `scikit-learn` of `Python` and corresponds to a grid of two-dimensional Gaussian blobs, see Fig. 7.9. For all experiments, the number of blobs is fixed to $n_{blobs} = 9$ *i.e.* and their mean are equidistant by a parameter $\theta \in \mathbb{R}^*$ that we make vary. Each blob is drawn from $\mathcal{N}(\mu, \sigma^2 \times \mathbb{I}_d)$, where μ has coordinates $\mu_i \in \{-\theta, 0, \theta\}$, for all $i \in \{1,2\}$, and $\sigma > 0$, $\sigma_Y = 1$.

(BL) *Location*. Both samples are drawn from the same distribution, but one is translated by $\varepsilon > 0$, with $\sigma_X^2 = \sigma_Y^2 = 1$ and $\theta = 3$.

(BS) *Scale*. Both samples are drawn from the same distribution, but one is rotated by an angle $\varepsilon \in \{\pi/4,\ \pi/6\}$, with $d \in \{2, 20\}$, $\sigma_X^2 \in \{1, 4\}$, $\sigma_Y^2 = 1$ and $\theta \in \{1, 3\}$.



a. Location blobs model ($\varepsilon = 1$) b. Rotation blobs model ($\varepsilon = \pi/6$)

Figure 7.9. Blobs datasets for the location model (a) and the rotation model (b), the sample drawn from G (blue) and from H (green), $n = m = 1000$, $\theta = 5$.

The framework (S1) is from Section 7.1, (S2), (S3) inspired from Deb and Sen (2019).

7.2.3 Results and discussion

This section discusses the numerical results obtained for the series of two-sample R-tests when comparing the empirical type-I error and power to state-of-the-art nonparametric tests. We detail the differences depending on the choice of the bipartite ranking algorithm. In particular, we concentrate on the ability to reject $\mathscr{H}_0$ for small deviations from it, while controlling low type-I error for challenging probabilistic models. We gather in tables these indicators for all designs when using the score-generating function $\phi = Id$, *i.e.*, corresponding to the MWW test statistic. Graphs are additionally plotted for other ϕ, in particular this of the RTB model with various proportion values $u_0 \in \{0.6, 0.7, 0.8, 0.9\}$. In this line, the optimal statistic(s) (in bold) is chosen to minimize the type-I error while maximizing the power for the smallest value of ε. All experiments are run using Python.

For all experiments, we performed the two-stage procedure detailed in Fig. 6.2 (Chap. 6) at fixed risk $\alpha = 0.05$ and sample sizes. We made the dimension d of the feature space and the discrepancy parameter ε vary. The more ε increases, the more the two underlying distributions are 'dissimilar'. First, all tests show increasing empirical power with the increase of ε, for all experiments except for the robust models. However, due to the sample sizes and the required data-split for all procedures,

we observe an instability in the results for some experiments despite a large number of Monte Carlo samplings (see *e.g.* MMD and Energy in Fig. 7.10, and when comparing their values in 7.1 and 7.1). We sequentially review the results for each category of the probabilistic model.

First and most importantly, the location models are under scrutiny. On the one hand, these are the most accurate probabilistic models for ranking algorithms. Precisely, *Step 1.* aims to learn the optimal scoring function by minimizing the bipartite ranking loss. It was extensively shown and discussed (Chap. 2 sec. 2.1, also Chap. 5) that only for the location model, the state-of-the-art ranking algorithms are designed to optimize the exact risk, either with the binary or surrogate losses. On the other hand, univariate rank statistics are known to be UMP for the location model; hence we expect similar behavior in the multivariate setting. Overall, ranking-based tests show a lower minimax separation rate than the three comparison tests. In particular, methods based on Neural Nets and those estimating the likelihood ratio (`SVR, LR`) reject the null for very small ε (Fig. 7.10). While both `RankSVM` and boosting methods do not falsely reject the null, they have similar performance to metric-based tests (Fig. 7.11, 7.12). In addition, Fig. 7.13 plots empirical ROC curves of the prediction obtained by the bipartite ranking algorithms for the (L1-) model to compare with the oracle curve, defined by ROC^*.

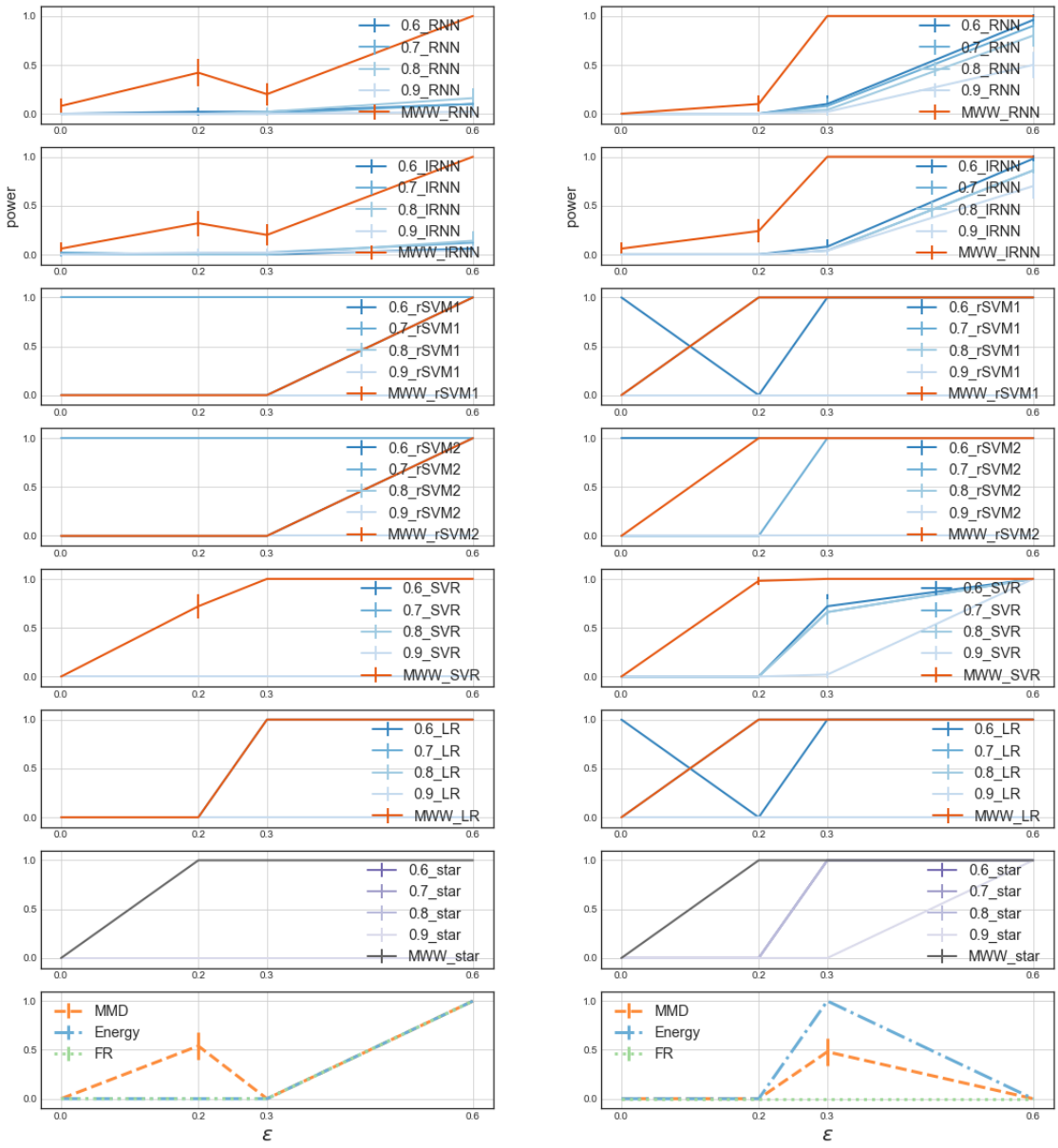
For the scale model (S1), we let the high dimension $d \in \{50, 100\}$. MMD and Energy tests have high statistical errors. On the contrary, `(Lambda)RankNN` have very small type-I error and an increase of the power for $d = 50$. For $d = 100$, only RTB and the Oracle tests are of interest. For both (S2) and (S3), `(Lambda)RankNN` models have a smaller ρ overall, while having higher power for greater dimension d (see Fig. 7.15). While FR has empirical fuzzy behavior, MMD (except for (S2) with $d = 20$) and Energy have lower testing performance. Lastly, boosting and SVM based tests do not converge in this setting, their empirical power is very low, hence we do not present their results.

The blobs dataset is of interest in this context for testing mixture types of models. In particular, the transformation between the two samples can be greater than each blob's variance. We compare the results for two different variances (σ) and means (θ). The empirical results discussed in Chwialkowski et al. (2015); Gretton et al. (2012c) showed low power for MMD and in particular if the kernel is not well-chosen (Gretton et al. (2012c)). Chwialkowski et al. (2015) later investigated it with a block method. This indicates that the implemented optimization of the kernel bandwidth circumvents, to a certain extent, this issue and more broadly, that the results obtained in this section are optimal in that sense. Regarding (BL), all ranking-based tests show high power, see Fig. 7.16 and Fig. 7.17. However, for (BS), SVM-based tests do not reject the null, even for large discrepancy parameter ε.

To conclude, the choice of score-generating function impacts the performance of the R-statistic. Of course, it was expected that $\phi = Id$ would yield better empirical results as it recovers the loss of the bipartite ranking algorithms, as fairly discussed. However, testing in *Step 2.* with ϕ_{rtb}, shows interesting results, in particular with `(Lambda)RankNN` models, see *e.g.* Fig. 7.10 ($d = 6$) and 7.15. Therefore we only can hope for improvements following the last Section 7.1 and as briefly detailed in Remark 10.

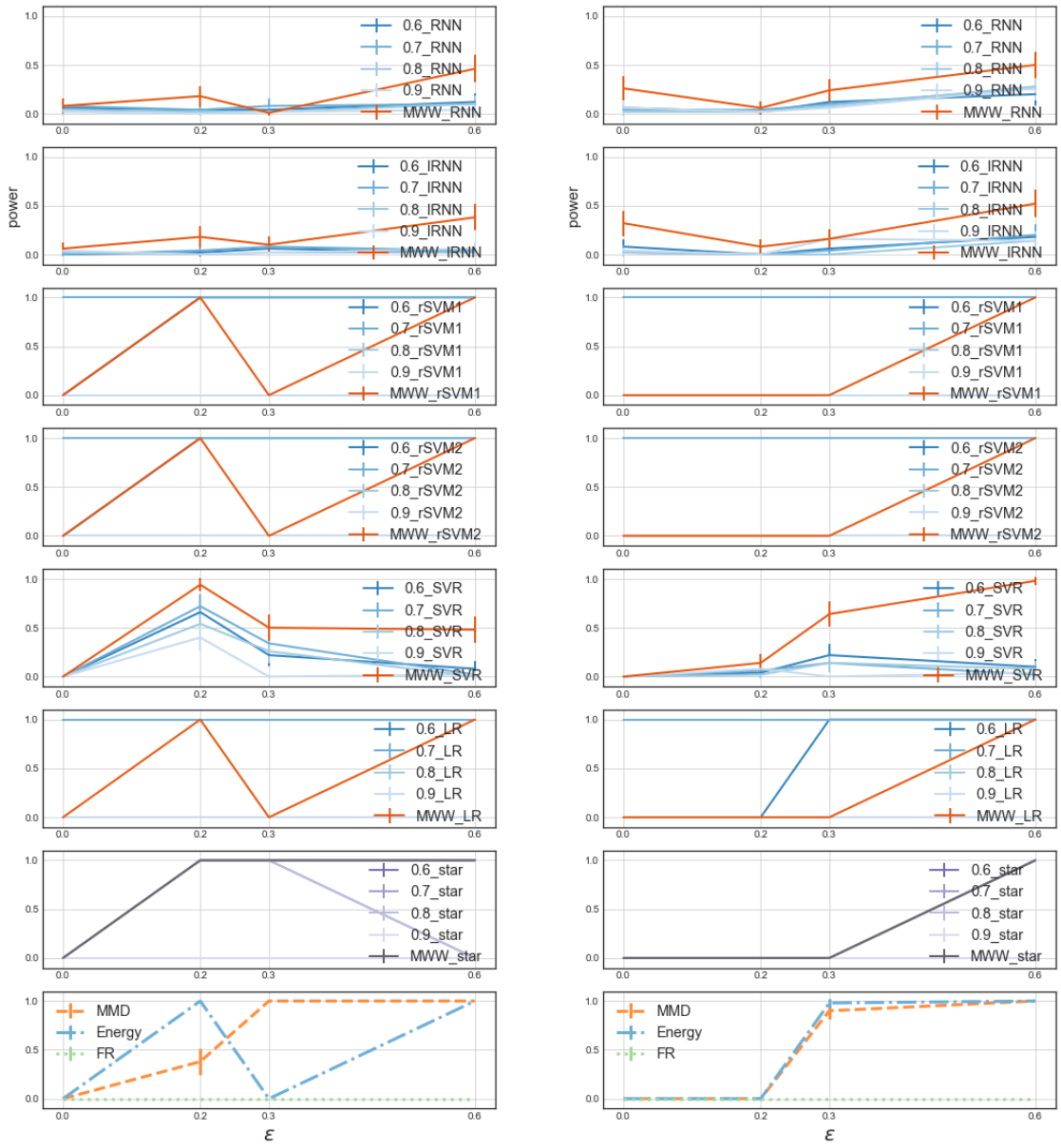
Method	**Type-I error**	**Power** ($d=4$)			**Type-I error**	**Power** ($d=6$)		
(L1-)	($d=4$)	$\varepsilon=0.2$	$\varepsilon=0.3$	$\varepsilon=0.6$	($d=6$)	$\varepsilon=0.2$	$\varepsilon=0.3$	$\varepsilon=0.6$
RNN	0.08 ($\pm$0.075)	0.42 ($\pm$0.137)	0.20 ($\pm$0.111)	**1.0**	**0.0**	0.1 ($\pm$0.083)	**1.0**	**1.0**
lRNN	0.06 ($\pm$0.066)	0.32 ($\pm$0.129)	0.20 ($\pm$0.111)	**1.0**	0.06 ($\pm$0.066)	0.24 ($\pm$0.112)	**1.0**	**1.0**
rSVM1	1.0	1.0	1.0	1.0	**0.0**	**1.0**	**1.0**	**1.0**
rSVM2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
rSVR	**0.0**	**0.72 ($\pm$0.124)**	**1.0**	**1.0**	**0.0**	0.98 ($\pm$0.039)	**1.0**	**1.0**
rLR	**0.0**	0.0	**1.0**	**1.0**	**0.0**	**1.0**	**1.0**	**1.0**
Oracle	**0.0**	**1.0**	**1.0**	**1.0**	**0.0**	**1.0**	**1.0**	**1.0**
MMD	**0.0**	**0.54 ($\pm$0.140)**	0.0	**1.0**	**0.0**	0.0	0.48 ($\pm$0.138)	0.0
Energy	**0.0**	0.0	0.0	**1.0**	**0.0**	0.0	**1.0**	0.0
FR	**0.0**	0.0	0.0	**1.0**	**0.0**	0.0	0.0	0.0
Method	**Type-I error**	**Power** ($d=4$)			**Type-I error**	**Power** ($d=6$)		
(L1+)	($d=4$)	$\varepsilon=0.2$	$\varepsilon=0.3$	$\varepsilon=0.6$	($d=6$)	$\varepsilon=0.2$	$\varepsilon=0.3$	$\varepsilon=0.6$
RNN	0.08 ($\pm$0.075)	0.18 ($\pm$0.106)	0.1 ($\pm$0.028)	0.46($\pm$0.138)	0.26 ($\pm$0.122)	0.06 ($\pm$0.066)	0.24 ($\pm$0.118)	0.5 ($\pm$0.139)
lRNN	0.06 ($\pm$0.066)	0.18 ($\pm$0.106)	0.1 ($\pm$0.028)	0.38 ($\pm$0.135)	0.32 ($\pm$0.129)	0.08 ($\pm$0.075)	0.16 ($\pm$0.102)	0.52 ($\pm$0.138)
rSVM1	**0.0**	**1.0**	0.0	**1.0**	**0.0**	0.0	0.0	**1.0**
rSVM2	**0.0**	**1.0**	0.0	**1.0**	0.0	0.0	0.0	**1.0**
rSVR	**0.0**	0.94 ($\pm$0.066)	0.50 ($\pm$0.139)	0.48 ($\pm$0.139)	**0.0**	0.14 ($\pm$0.096)	0.64($\pm$0.133)	**0.98 ($\pm$0.039)**
rLR	**0.0**	0.0	**1.0**	0.0	**0.0**	0.0	0.0	**1.0**
Oracle	**0.0**	**1.0**	**1.0**	**1.0**	**0.0**	0.0	0.0	**1.0**
MMD	**0.0**	0.0	0.38($\pm$0.135)	**1.0**	0.0	0.0	0.9 ($\pm$0.083)	**1.0**
Energy	**0.0**	0.0	1.0	0.0	0.0	0.0	**0.98 ($\pm$0.039)**	**1.0**
FR	**0.0**	0.0	0.0	0.0	**0.0**	0.0	0.0	0.0

Table 7.1. Estimation of type-I error and power for the location models (L1-) and (L1+) with $n,m = 500,500$ and $d \in \{4,6\}$, $\pm$ their standard deviation at 95%. For ranking methods, only the results associated to MWW test are presented. Bold estimates represent the ones that among all algorithms minimize the type-I error, and maximize the power. The algorithm having 'best' empirical results is in bold.



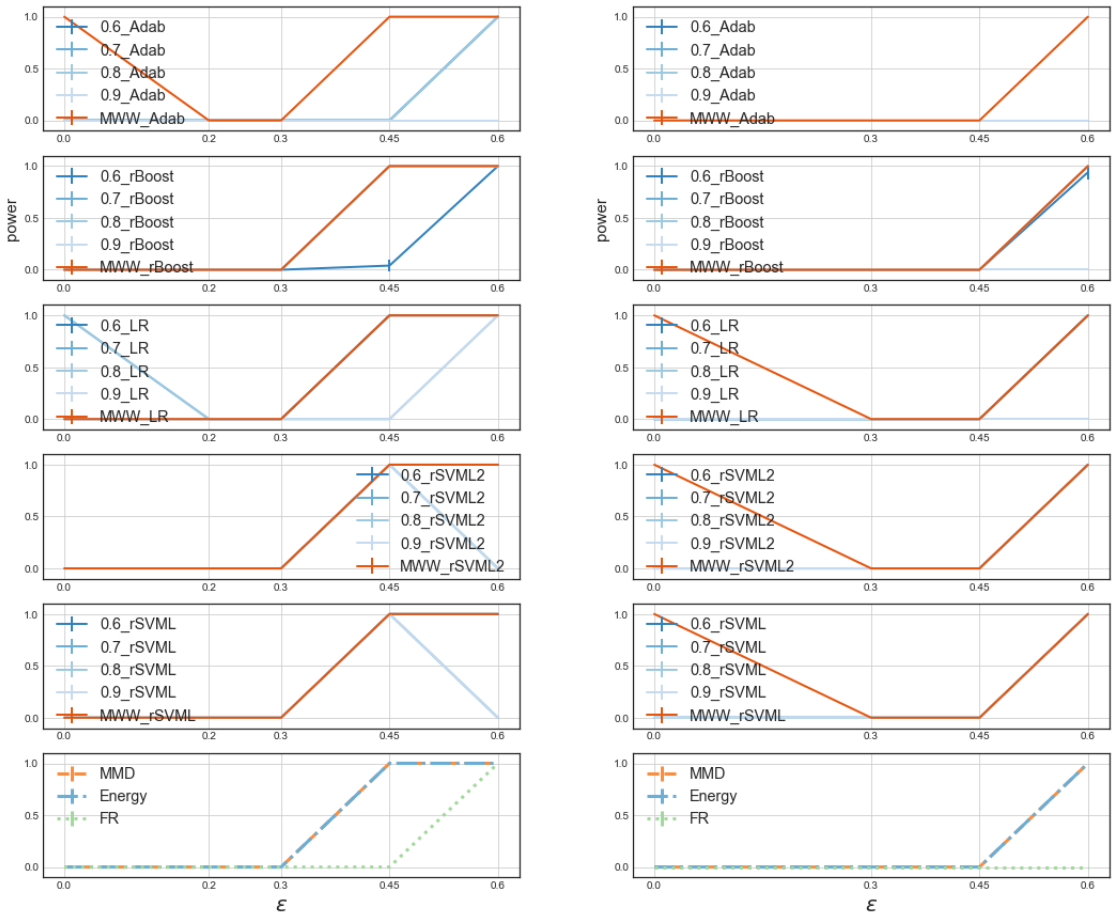
a. (L1-), $d = 4$ b. (L1-), $d = 6$

Figure 7.10. Estimation of type-I error and power for the model (L1-) with $n, m = 500, 500$, $d = 4$ (a) and $d = 6$ (b), $\pm$ their standard deviation at 95%. Results for each algorithm are gathered in a same plot, where RTB tests with $u_0 \in \{0.6, 0.7, 0.8, 0.9\}$ are estimated.



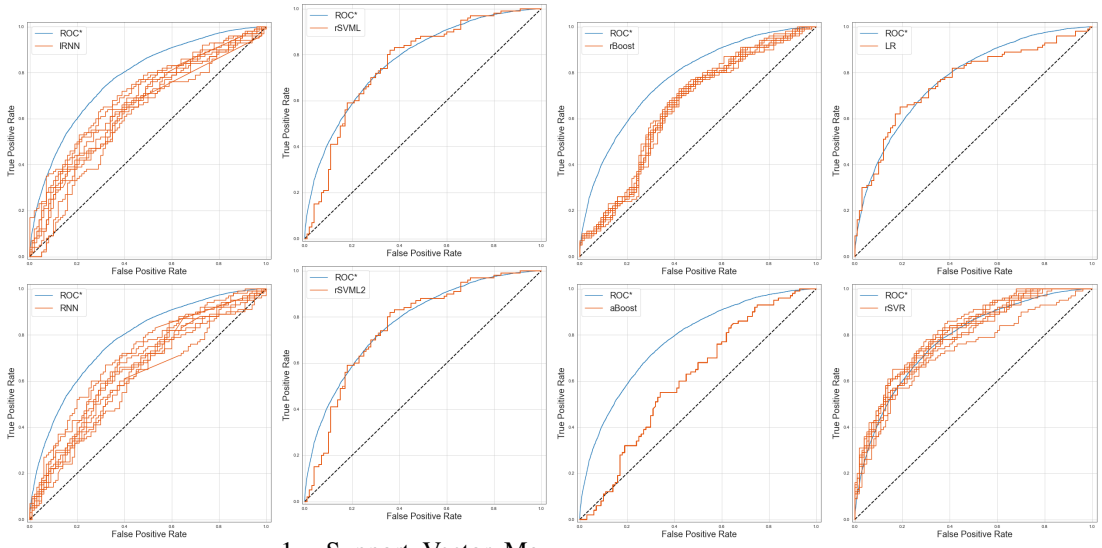
a. (L1+), $d = 4$ b. (L1+), $d = 6$

Figure 7.11. Estimation of type-I error and power for the model (L1+) with $n, m = 500, 500$, $d = 4$ (a) and $d = 6$ (b), $\pm$ their standard deviation at 95%. Results for each algorithm are gathered in a same plot, where RTB tests with $u_0 \in \{0.6, 0.7, 0.8, 0.9\}$ are estimated.



a. (L1+), $d = 4$ b. (L1+), $d = 6$

Figure 7.12. Estimation of type-I error and power for (L1+) model with $n, m = 500, 500$, and with $d = 4$ (a) and $d = 6$ (b), $\pm$ their standard deviation at 95%. Results for each algorithm are gathered in a same plot, where RTB tests with $u_0 \in \{0.6, 0.7, 0.8, 0.9\}$ are estimated.

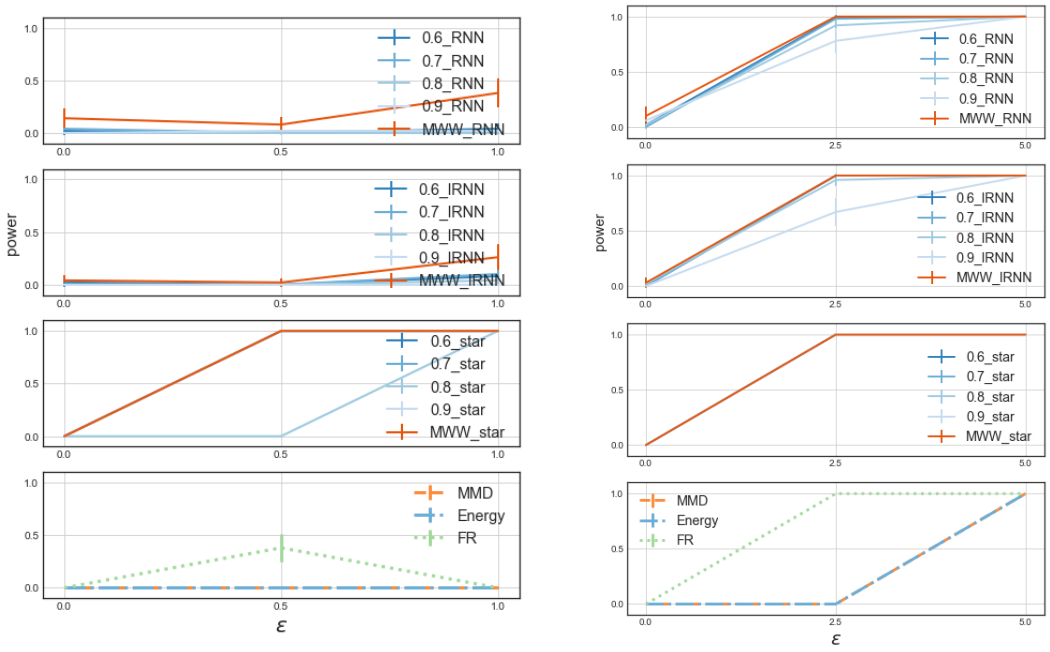


1. Neural Nets | 1. Support Vector Machine | 1. Boosting | 1. Binary CPE

Figure 7.13. Empirical ROC curves of the predictions of scores for the observations drawn from model (L1-) with $\varepsilon = 0.20$ and $d = 6$. The ranking algorithms are `RankNN` and `LambdaRankNN` (1.), `RankSVM` with logistic regression, L_1 and L_2 loss (2.), `RankBoost` and `AdaBoost` (3.). Logistic regression and linear support vector regression (4.). The blue curve corresponds to ROC*.

Method (S1)	**Type-I error** ($d = 50$)	**Power** ($d = 50$) $\varepsilon = 0.5$	$\varepsilon = 1.0$	**Type-I error** ($d = 100$)	**Power** ($d = 100$) $\varepsilon = 2.5$	$\varepsilon = 5.0$
RNN	0.14 ($\pm$0.096)	0.08 ($\pm$0.075)	**0.38 ($\pm$0.135)**	0.1 ($\pm$0.083)	1.0	1.0
lRNN	**0.04 ($\pm$0.054)**	0.02 ($\pm$0.039)	0.26 ($\pm$0.122)	**0.03 ($\pm$0.047)**	1.0	1.0
Oracle	**0.0**	**1.0**	**1.0**	0.0	**1.0**	**1.0**
MMD	**0.0**	0.0	0.0	**0.0**	0.0	**1.0**
Energy	**0.0**	0.0	0.0	**0.0**	0.0	**1.0**
FR	**0.0**	**0.38 ($\pm$0.146)**	0.0	**0.0**	**1.0**	**1.0**
Method (S2)	**Type-I error** ($d = 3$)	**Power** ($d = 3$) $\varepsilon = 0.2$	$\varepsilon = 0.4$	**Type-I error** ($d = 20$)	**Power** ($d = 20$) $\varepsilon = 0.2$	$\varepsilon = 0.4$
RNN	0.08 ($\pm$0.075)	**0.32 ($\pm$0.129)**	0.98 ($\pm$0.039)	0.18 ($\pm$0.106)	**0.82 ($\pm$0.106)**	**1.0**
lRNN	**0.06 ($\pm$0.066)**	0.18 ($\pm$0.106)	**1.0**	**0.10 ($\pm$0.083)**	0.76 ($\pm$0.118)	**1.0**
Oracle	**0.0**	**1.0**	**1.0**	0.0	**1.0**	**1.0**
MMD	**0.0**	0.0	**1.0**	**0.0**	**1.0**	**1.0**
Energy	**0.0**	0.0	**0.78 ($\pm$0.115)**	**0.0**	0.0	**1.0**
FR	1.0	0.0	**1.0**	1.0	**1.0**	**1.0**
Method (S3)	**Type-I error** ($d = 3$)	**Power** ($d = 3$) $\varepsilon = 0.2$	$\varepsilon = 0.4$	**Type-I error** ($d = 20$)	**Power** ($d = 20$) $\varepsilon = 0.1$	$\varepsilon = 0.2$
RNN	0.04 ($\pm$0.054)	**0.42 ($\pm$0.137)**	0.98 ($\pm$0.039)	**0.02($\pm$0.039)**	**0.30($\pm$0.127)**	0.80($\pm$0.111)
lRNN	0.10 ($\pm$0.083)	0.34 ($\pm$0.131)	**1.0**	0.04 ($\pm$0.054)	0.22 ($\pm$0.115)	**0.90 ($\pm$0.138)**
Oracle	**0.0**	**1.0**	**1.0**	**0.0**	**1.0**	**1.0**
MMD	0.58 ($\pm$0.137)	0.0	**1.0**	0.0	0.0	**1.0**
Energy	0.10 ($\pm$0.083)	0.0	**1.0**	0.0	0.0	0.48 ($\pm$0.138)
FR	**0.0**	0.0	**1.0**	0.0	0.0	0.02 ($\pm$0.039)

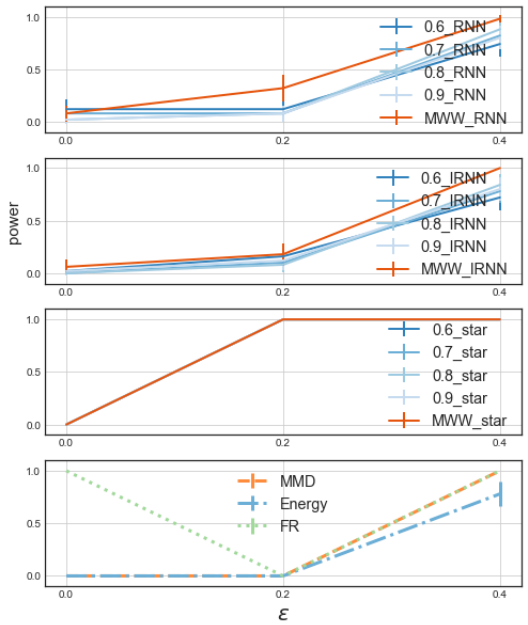
Table 7.2. Estimation of type-I error and power for the scale models (S1), (S2) and (S3) with $n, m = 500, 500$, and $d \in \{50, 100\}$ (for S1) $d \in \{3, 20\}$ (for S2, S3), $\pm$ their standard deviation at 95%. For ranking methods, only the results associated to MWW test are presented. Bold estimates represent the ones that among all algorithms minimize the type-I error, and maximize the power. The algorithm having 'best' empirical results is in bold.



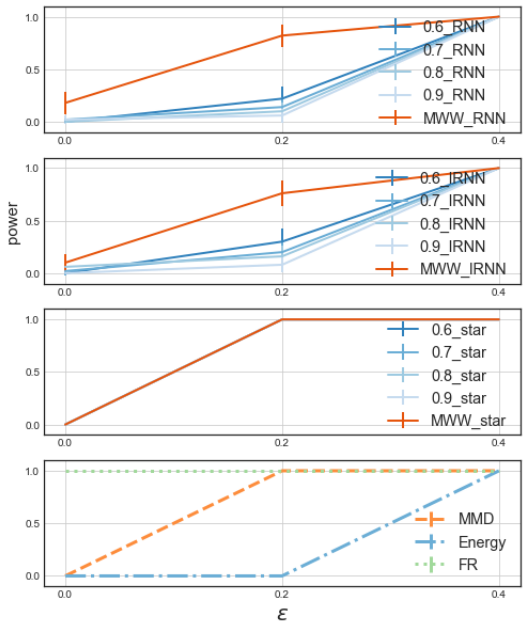
a. (S1), $d = 50$

b. (S1), $d = 100$

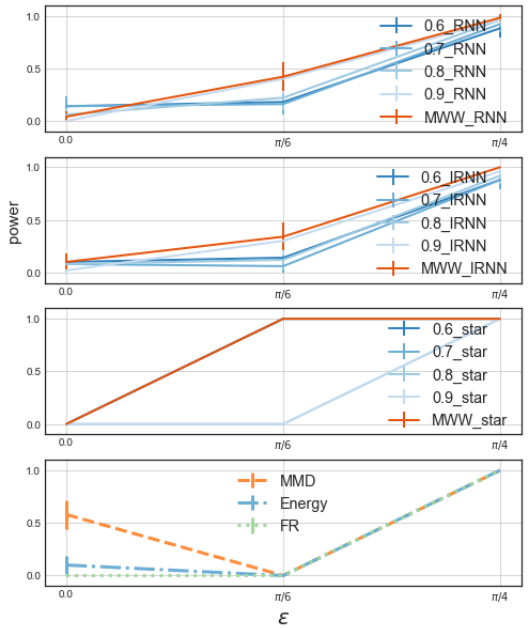
Figure 7.14. Estimation of type-I error and power for the scale model (S1) with $n, m = 500, 500$ and with $d = 50$ (a) and $d = 100$ (b), $\pm$ their standard deviation at 95%. Results for each algorithm are gathered in a same plot, where RTB tests with $u_0 \in \{0.6, 0.7, 0.8, 0.9\}$ are estimated.



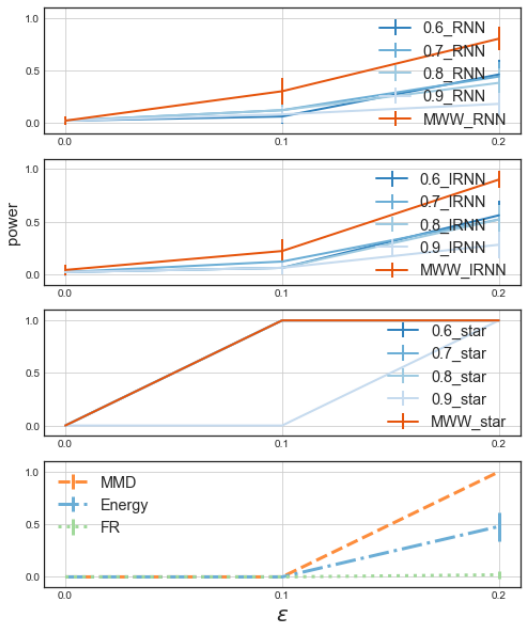
a. (S2), $d = 3$



b. (S2), $d = 20$



c. (S3), $d = 3$

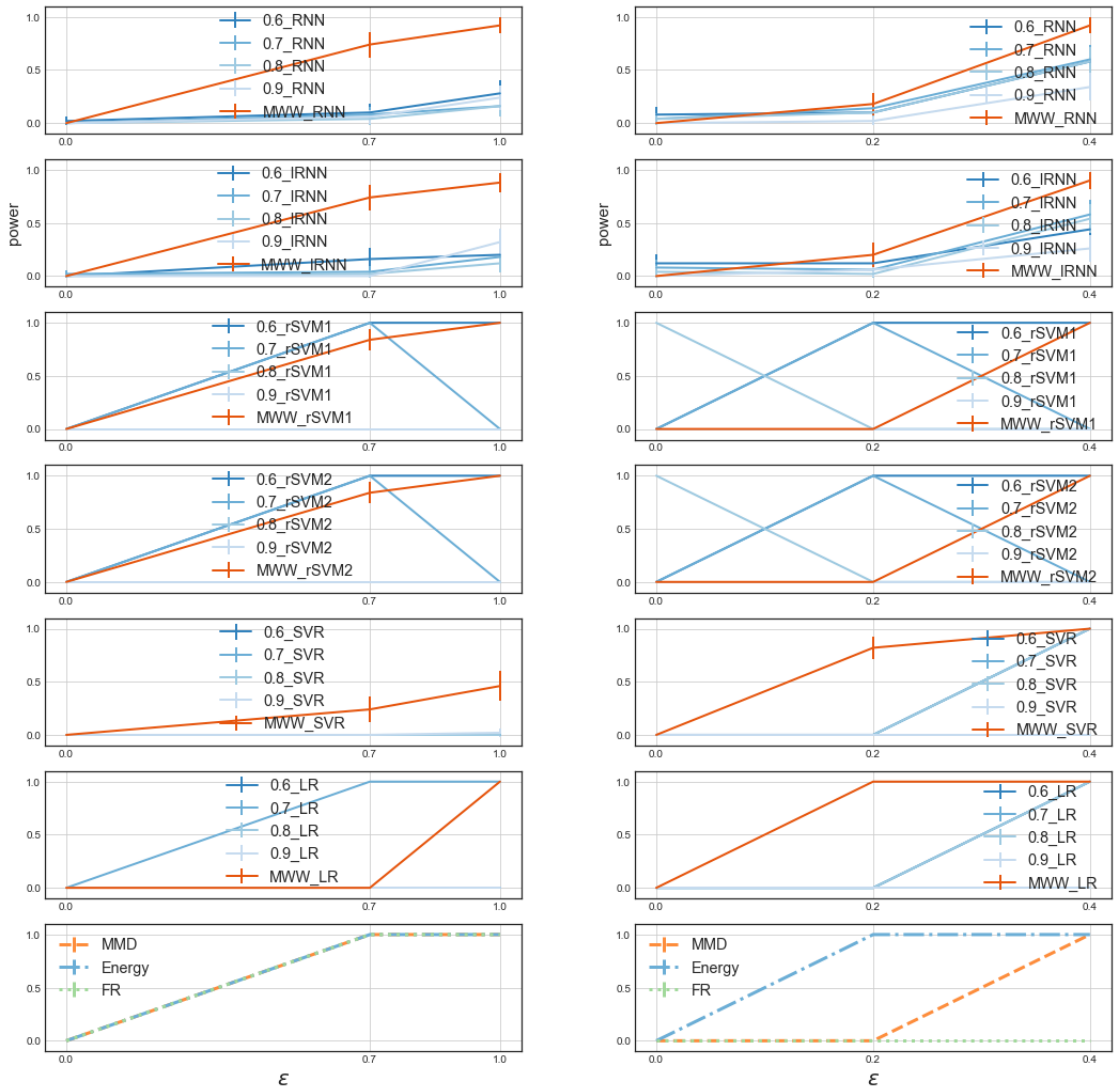


d. (S3), $d = 20$

Figure 7.15. Estimation of type-I error and power for the scale models (S2) (a,b) and (S3) (c,d) with $n, m = 500, 500$ and with $d = 3$ (a,c) and $d = 20$ (b,d), $\pm$ their standard deviation at 95%. Results for each algorithm are gathered in a same plot, where RTB tests with $u_0 \in \{0.6, 0.7, 0.8, 0.9\}$ are estimated.

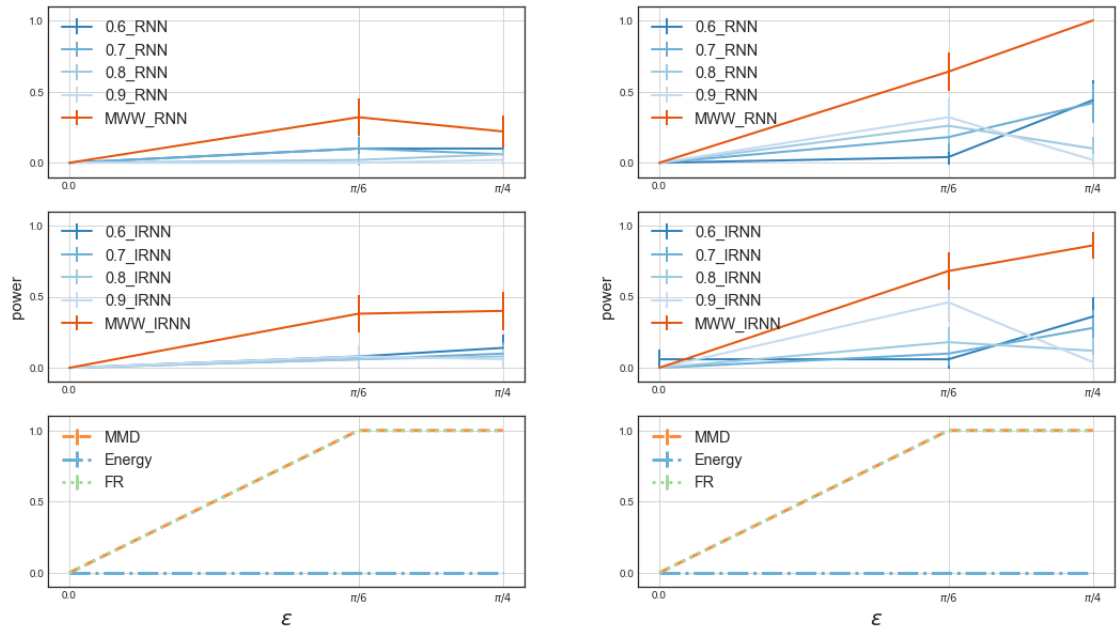
Method (BL)	**Type-I error** ($d=2,\ \theta=3$)	**Power** ($\theta=3$) $\varepsilon=0.7$	$\varepsilon=1.0$	**Type-I error** ($d=2,\ \theta=1$)	**Power** $\theta=1$ $\varepsilon=0.2$	$\varepsilon=0.4$
RNN	**0.0**	0.74 ($\pm$ 0.122)	0.92 ($\pm$0.075)	**0.0**	0.18 ($\pm$ 0.106)	0.92 ($\pm$ 0.075)
lRNN	**0.0**	0.74 ($\pm$ 0.122)	0.88 ($\pm$0.090)	**0.0**	0.2 ($\pm$ 0.111)	0.9 ($\pm$ 0.083)
rSVM1	**0.0**	**0.84 ($\pm$0.102)**	**1.0**	**0.0**	0.0	**1.0**
rSVM2	**0.0**	**0.84 ($\pm$0.102)**	**1.0**	**0.0**	0.0	**1.0**
rSVR	**0.0**	0.24 ($\pm$0.118)	0.46 ($\pm$0.138)	**0.0**	0.82 ($\pm$ 0.106)	**1.0**
rLR	**0.0**	0.0	1.0	**0.0**	**1.0**	**1.0**
MMD	**0.0**	**1.0**	**1.0**	**0.0**	0.0	**1.0**
Energy	**0.0**	**1.0**	**1.0**	**0.0**	**1.0**	**1.0**
FR	**0.0**	**1.0**	**1.0**	**0.0**	0.0	0.0
Method (BS)	**Type-I error** ($d=2,\ \theta=3$)	**Power** ($d=3$) $\varepsilon=\pi/6$	$\varepsilon=\pi/4$	**Type-I error** ($d=20,\ \theta=3$)	**Power** ($d=20$) $\varepsilon=\pi/6$	$\varepsilon=\pi/4$
RNN	**0.0**	0.32 ($\pm$0.129)	0.2 ($\pm$0.115)	0.0	0.64 ($\pm$0.133)	1.0
lRNN	**0.0**	0.38 ($\pm$0.135)	0.40 ($\pm$0.136)	0.0	0.68 ($\pm$0.129)	0.86 ($\pm$0.096)
MMD	**0.0**	**1.0**	**1.0**	**0.0**	**1.0**	**1.0**
Energy	**0.0**	0.0	0.0	**0.0**	0.0	0.0
FR	**0.0**	**1.0**	**1.0**	**0.0**	**1.0**	**1.0**

Table 7.3. Estimation of type-I error and power for the blobs models with $n,m = 500,500$, $d \in \{2,20\}$ and $\theta \in \{1,3\}$, $\pm$ their standard deviation at 95%. For ranking methods, only the results associated to MWW test are presented. Bold estimates represent the ones that among all algorithms minimize the type-I error, and maximize the power. The algorithm having 'best' empirical results is in bold.



a. (BL), $\theta = 3$ b. (BL), $\theta = 1$

Figure 7.16. Estimation of type-I error and power for the blobs location model (BL) with $n, m = 500, 500$, $d = 2$ and with $\theta = 3$ (a) and $\theta = 1$ (b), $\pm$ their standard deviation at 95%. Results for each algorithm are gathered in a same plot, where RTB tests with $u_0 \in \{0.6, 0.7, 0.8, 0.9\}$ are estimated.



a. (RS), $d = 2, \theta = 3, \sigma = 1$

b. (RS), $d = 20, \theta = 3, \sigma = 1$

c. (RS), $d = 2, \theta = 1, \sigma_X = 4$

d. (RS), $d = 2, \theta = 1, \sigma = 1$

Figure 7.17. Estimation of type-I error and power for the blobs rotation model (BS) with $n, m = 500, 500$ and with $d = 2, \theta = 3$ (a), $d = 20, \theta = 3$ (b) and $d = 2, \theta = 1, \sigma = 4$ (c), $d = 2, \theta = 1, \sigma = 1$ (d), $\pm$ their standard deviation at 95%. Results for each algorithm are gathered in a same plot, where RTB tests with $u_0 \in \{0.6, 0.7, 0.8, 0.9\}$ are estimated.

7.3 Conclusion

We exposed the empirical results of the procedures introduced in Chapter 5 for bipartite ranking and 6 for the two-sample problem. First, the deterministic gradient ascent algorithm applied to bipartite ranking shows that maximizing the proposed smoothed empirical W_ϕ-ranking criteria outputs scoring functions that converge to the oracle model in the ROC space. In particular, different score-generating functions ϕ reveal various characteristics of the underlying distributions. We highlight that concentrating the learning on the 'best' instances yields a better recovery of the beginning of the oracle

ROC curve. Then, the two-stage ranking procedure showed interesting properties of state-of-the-art ranking algorithms for the two-sample problem. Indeed, a series of probabilistic models were tested. Generally, they showed comparable empirical type-I and type-II errors to classical tests: Maximum Mean Discrepancy statistic, Energy-based statistic, and Wald-Wolfowitz runs statistic. Open access online codes also accompany all the experiments.

7.4 Appendix

Learning-to-rank algorithms. We detail how we modified the learning-to-rank algorithms in order to fit the test statistic introduced. First, we fix a same index of query for all observations and fix the score to 1 (*resp.* to 0) for the first sample drawn from G (*resp.* H). The three learning-to-rank algorithms construct all possible couples of observations from the pooled sample. Here we modify this step such that for all couples the first instance is drawn from G to obtain the sequence: $(\mathbf{X}_i,\mathbf{X}_1),\ldots,(\mathbf{X}_i,\mathbf{X}_i),\ldots,(\mathbf{X}_i,\mathbf{X}_n)$ and $(\mathbf{X}_i,\mathbf{Y}_1),\ldots,(\mathbf{X}_i,\mathbf{Y}_m)$, for all $i \leq n$. Besides these points, the structure of both algorithms is kept identical, note incidently that a smaller number of gradients are computed in the Neural Net structure. The modified algorithms are detailed in Algorithm 4.

Algorithm 4: Modified Pairwise Learning-to-Rank algorithm procedure.

Data: Dataset of independent *i.i.d.* samples $\{\mathbf{X}_i\}_{i\leq n}$ and $\{\mathbf{Y}_j\}_{j\leq m}$.
Input: Learning-to-Rank algorithm RANK, Score-generating functions ϕ, parameters of the NN (expliciter).
Result: Ranking model

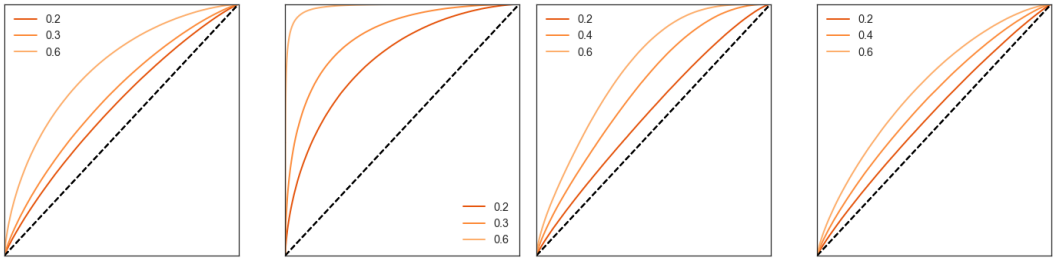
```
Set q = (1, ..., 1) of size n+m
Set score = (1, ..., 1, 0, ..., 0) of size n+m
for i = 1, ..., n do
    for k = 1, ..., n+m do
        store the couple (X_i, Z_k), where Z_k = X_k if k ≤ n, Z_k = Y_{k-n} if k ≥ n
    end
end
Build the model RANK((X_i, Z_k)_{i≤n, k≤n+m})
```

Exact parameters for the location models.

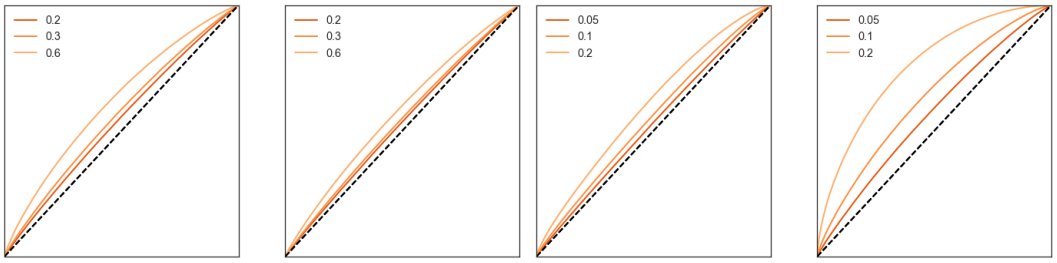
(L1−) for $d = 4$, the diagonals of Σ are $(2,6,1,5)$, $(-1,0,0)$, $(-1,0)$, (-1); $d = 6$, the matrix is extended in a similar way with the main diagonal equal to $(2,6,1,5,4,3)$

(L1+) for $d = 4$, the diagonals of Σ are $(6,4,5,3)$, $(-2,4,2)$, $(-3,0)$, (-2); $d = 6$, the matrix diagonals are equal to $(6,5,5,3,2,3)$, $(-2,4,2,1,2)$, $(-3,0,0,1)$, $(-2,1,1)$, $(-3,2)$, (-2).

ROC* curves for the location and the scale models.



(a) (L1-), left: $d = 4$, right: $d = 6$ **(c)** (S1), left: $d = 3$, right: $d = 20$



(b) (L1+), left: $d = 4$, right: $d = 6$ **(d)** (S2), left: $d = 4$, right: $d = 20$

Figure 7.18. True ROC curves (ROC*) for the location and the scale models: (L1+),(L1-),(S1),(S2), depending on the discrepancy parameter $\varepsilon \in \{0.2, 0.3, 0.6\}$ for (L1+) and (L1-), $\varepsilon \in \{0.2, 0.4, 0.6\}$ for (S1) and $\varepsilon \in \{0.05, 0.1, 0.2\}$ for (S2).

Part III

Applications

8 | Learning to Rank Anomalies with Two-sample Linear R-statistics

Abstract. The ability to collect and store ever more massive databases has been accompanied by the need to process them efficiently. In many cases, most observations have the same behavior, while a probable small proportion of these observations are abnormal. Detecting the latter, defined as outliers, is one of the major challenges for machine learning applications (*e.g.* in fraud detection or in predictive maintenance). In this chapter and following Section 2.3 (Chapter 2), we propose a methodology addressing the problem of outlier detection, by learning a data-driven scoring function defined on the feature space which reflects the degree of abnormality of the observations. This scoring function is learnt through a well-designed binary classification problem whose empirical criterion takes the form of a two-sample linear rank statistics on which theoretical results are available. We illustrate our methodology with preliminary encouraging numerical experiments.

Contents

8.1 Introduction

The problem of ranking multivariate data by degree of abnormality, referred to as *anomaly ranking*, is of central importance for a wide variety of applications (*e.g.* fraud detection, fleet monitoring, predictive maintenance). In the standard setup, the 'normal' behavior of the system under study (in the sense of 'not abnormal', without any link to the Gaussian distribution) is described by the (unknown) distribution $F(dx)$ of a generic *r.v.* X, valued in $\mathbb{R}^d$. The goal pursued is to build a scoring function $s : \mathbb{R}^d \to \mathbb{R}_+ \cup \{+\infty\}$ that ranks any observations $x_1, \ldots, x_n$ nearly in the same order as any increasing transform of the density f would do. Ideally, the smaller the score $s(x)$ of an observation x in $\mathbb{R}^d$, the more abnormal it should be considered. In Clémençon and Thomas (2018), a functional criterion, namely a Probability-Measure plot referred to as the *Mass-Volume* curve (the MV curve in abbreviated form), has been proposed to evaluate the anomaly ranking performance of any scoring rule $s(x)$. This performance measure can be viewed as the unsupervised version of the *Receiver Operating Characteristic* (ROC) curve, the gold standard measure to evaluate the accuracy of scoring functions in the bipartite ranking context, see *e.g.* Clémençon and Vayatis (2009b). Beyond this approach, let us highlight that the problem of anomaly detection has also been studied *via* various other modelings. For instance, the works of Bergman and Hoshen (2020) and Steinwart et al. (2005) are based on classification methods, while Liu et al. (2008) build on peeling, Breunig et al. (2000) on local averaging criteria, Frery et al. (2017) on ranking and Schölkopf et al. (2001) on plug-in techniques.

In this chapter, we propose a novel two-stage method for detecting and ranking abnormal instances, by means of scalar criteria summarizing the MV curve and extending the area under its curve, when $F(dx)$ has compact support. Briefly, starting from a sample of observations $X_1, \ldots, X_n$, we artificially generate an independent second sample $U_1, \ldots, U_m$ that is used as a proxy for outliers. For theoretical reasons explained in the chapter, the agnostic choice consists in sampling the U_i's *i.i.d.* from the uniform law on a subset of $\mathbb{R}^d$, which $F(dx)$'s support is supposedly included in. We then learn to discriminate the X_i's from the U_i's thanks to a scoring function that maximizes two-sample empirical counterparts of the aforementioned criteria, that are in particular robust to imbalanced datasets. The resulting scoring function allows to rank the X_i's by degree of abnormality. This novel class of criteria is based on theoretical guarantees provided by Clémençon et al. (2021) on general classes of two-sample linear rank processes, that incidentally circumvent the difficulty of optimizing the functional MV criterion. Beyond the classical results of statistical learning theory for these processes, Clémençon et al. (2021) obtain theoretical generalization guarantees for their empirical optimizers. The numerical results performed at the end of the chapter also provide strong empirical evidence of the relevance of the approach promoted here.

The chapter is structured as follows. In section 8.2, the formulation of the (unsupervised) anomaly ranking problem is recalled at length, together with the concept of MV curve. In section 10.2, the anomaly ranking performance criteria proposed are introduced and their statistical estimation is discussed. Optimization of the statistical counterparts of the criteria introduced to build accurate anomaly scoring functions is also put forward therein. Finally, the relevance of this approach is illustrated by numerical results in section 8.4.

8.2 Background and preliminaries

We start off with recalling the formulation of the (unsupervised) anomaly ranking problem and introducing notations that shall be used here and throughout. By λ is meant the Lebesgue measure on $\mathbb{R}^d$, by $\mathbb{I}\{\mathscr{E}\}$ the indicator function of any event $\mathscr{E}$, while the generalized inverse of any cumulative distribution function $K(t)$ on $\mathbb{R}$ is denoted by $K^{-1}(u) = \inf\{t \in \mathbb{R} : K(t) \geq u\}$. We consider a *r.v.* X valued in $\mathbb{R}^d$, $d \geq 1$, with distribution $F(dx) = f(x)\lambda(dx)$, modeling the 'normal' behavior of the system under study. The observations at disposal $X_1, \ldots, X_n$, with $n \geq 1$, are independent copies of X. Based on the X_i's our goal is to learn a ranking rule for deciding among two observations x and x' in $\mathbb{R}^d$ which one is more 'abnormal'. The simplest way of defining a preorder[1] on $\mathbb{R}^d$ consists in transporting the natural order on $\mathbb{R}_+ \cup \{+\infty\}$ onto it through a *scoring function*, *i.e.* a Borel measurable mapping $s : \mathbb{R}^d \to \mathbb{R}_+$: given two observations x and x' in $\mathbb{R}^d$, x is said to be more abnormal according to s than x' when $s(x) \leq s(x')$. The set of all anomaly scoring functions that are integrable with respect to Lebesgue measure is denoted by $\mathscr{S}$. The integrability condition is not restrictive since the preorder induced by any scoring function is invariant under strictly increasing transformation (*i.e.* the scoring function s and its transform $T \circ s$ define the same preorder on $\mathbb{R}^d$ provided that the Borel measurable transform $T : \mathrm{Im}(s) \to \mathbb{R}_+$ is strictly increasing on the image of the *r.v.* $s(X)$, denoted by $\mathrm{Im}(s)$). One wishes to build, from the 'normal' observations only, a scoring function s such that, ideally, the smaller $s(X)$, the more abnormal the observation X. The set of optimal scoring rules in $\mathscr{S}$ should be thus composed of strictly increasing transforms of the density function $f(x)$ that are integrable *w.r.t.* to λ, namely:

$$\mathscr{S}^* = \{T \circ f : T : \mathrm{Im}(f) \to \mathbb{R}_+ \text{ strictly increasing}, \int_{\mathbb{R}^d} T \circ f(x)\lambda(dx) < +\infty\} . \tag{8.2.1}$$

The technical assumptions listed below are required to define a criterion, whose optimal elements coincide with $\mathscr{S}^*$.

H$_1$ The *r.v.* $f(X)$ is continuous, *i.e.* $\forall c \in \mathbb{R}_+, \mathbb{P}\{f(X) = c\} = 0$.

H$_2$ The density function $f(x)$ is bounded: $||f||_\infty \overset{def}{=} \sup_{x \in \mathbb{R}^d} |f(x)| < +\infty$.

Measuring anomaly scoring accuracy - The MV curve. Consider an arbitrary scoring function $s \in \mathscr{S}$ and denoted by $\Omega_{s,t} = \{x \in \mathscr{X} : s(x) \geq t\}$, $t \geq 0$, its level sets. As s is λ-integrable, the measure $\lambda(\Omega_{s,t}) \leq (\int_{u \in \mathbb{R}_+} s(u)du)/t$ is finite for any $t > 0$. Introduced in Clémençon and Thomas (2018), a natural measure of the anomaly ranking performance of any scoring function candidate s is the Probability-Measure plot, referred to as the *Mass-Volume* (MV) curve

$$t > 0 \mapsto \Big(\mathbb{P}\{s(X) \geq t\}, \lambda(\{x \in \mathbb{R}^d : s(x) \geq t\})\Big) = (F(\Omega_{s,t}), \lambda(\Omega_{s,t})) . \tag{8.2.2}$$

[1] A preorder $\preccurlyeq$ on a set $\mathscr{Z}$ is a reflexive and transitive binary relation on $\mathscr{Z}$. It is said to be *total*, when either $z \preccurlyeq z'$ or else $z' \preccurlyeq z$ holds true, for all $(z, z') \in \mathscr{Z}^2$.

Connecting points corresponding to possible jumps, this parametric curve can be viewed as the plot of the continuous mapping $\mathrm{MV}_s : \alpha \in (0,1) \mapsto \mathrm{MV}_s(\alpha)$, starting at $(0,0)$ and reaching $(1, \lambda\left(\mathrm{supp}(F)\right))$ in the case where the support $\mathrm{supp}(F)$ of the distribution $F(dx)$ is compact, or having the vertical line '$\alpha = 1$' as an asymptote otherwise. A typical MV curve is depicted in Fig. 8.1.

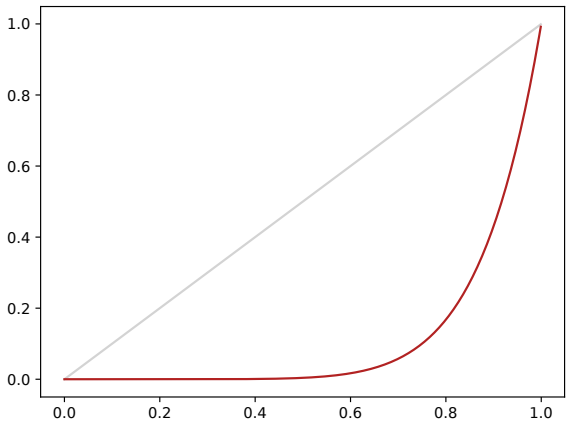


Figure 8.1. Typical MV curve in red (x-axis:volume, y-axis:mass). In gray, the diagonal $y = x$.

Let $\alpha \in (0,1)$. Denoting by $F_s(t)$ the cumulative distribution function of the *r.v.* $s(X)$, we have:

$$\mathrm{MV}_s(\alpha) = \lambda\left(\{x \in \mathbb{R}^d : \ s(x) \geq F_s^{-1}(1-\alpha)\}\right), \tag{8.2.3}$$

when $F_s \circ F_s^{-1}(\alpha) = \alpha$. This functional criterion is invariant by increasing transform and induces a partial order over the set $\mathscr{S}$. Let $(s_1, s_2) \in \mathscr{S}^2$, the ordering defined by s_1 is said to be more accurate than the one induced by s_2 when:

$$\forall \alpha \in (0,1), \ \ \mathrm{MV}_{s_1}(\alpha) \leq \mathrm{MV}_{s_2}(\alpha) \ .$$

As summarized by the result stated below, the MV curve criterion is adequate to measure the accuracy of scoring functions with respect to anomaly ranking. It reveals in particular that optimal scoring functions are those whose MV curve is minimum everywhere.

Proposition 79 (Clémençon and Thomas (2018))**.** *Let the assumptions* $\mathbf{H_1} - \mathbf{H_2}$ *be fulfilled. The elements of the class* $\mathscr{S}^*$ *have the same (convex)* MV *curve and provide the best possible preorder on* $\mathbb{R}^d$ *w.r.t. the* MV *curve criterion:*

$$\forall (s, \alpha) \in \mathscr{S} \times (0,1), \ \ \mathrm{MV}^*(\alpha) \leq \mathrm{MV}_s(\alpha) \ , \tag{8.2.4}$$

where $\mathrm{MV}^*(\alpha) = \mathrm{MV}_f(\alpha)$ *for all* $\alpha \in (0,1)$.

Equation (8.2.4) reveals that the lowest the MV curve (everywhere) of a scoring function $s(x)$, the closer the preorder defined by $s(x)$ is to that induced by $f(x)$. Favorable situations are those where the MV curve increases slowly and rises more rapidly when coming closer to the 'one' value: this correponds to the case where $F(dx)$ is much concentrated around its modes, $s(X)$ takes its highest values near the latter and its lowest values are located in the tail region of the distribution $F(dx)$. Incidentally, observe that the optimal curve MV^* somehow measures the spread of the distribution

$F(dx)$ in particular for large values of α *w.r.t.* extremal observations (*e.g.* a light tail behavior corresponds to the situation where $\mathrm{MV}^*(\alpha)$ increases rapidly when approaching 1), whereas it should be examined for small values of α when modes of the underlying distributions are investigated (a flat curve near 0 indicates a high degree of concentration of $F(dx)$ near its modes).

Statistical estimation. In practice, the MV curve of a scoring function $s \in \mathscr{S}$ is generally unknown, just like the distribution $F(dx)$, and it must be estimated. A natural empirical counterpart can be obtained by plotting the stepwise graph of the mapping:

$$\widehat{\mathrm{MV}}_s(\alpha) : \alpha \in (0,1) \mapsto \lambda\left(\left\{x \in \mathbb{R}^d : \ s(x) \geq \widehat{F}_{s,n}^{-1}(1-\alpha)\right\}\right) , \tag{8.2.5}$$

where $\widehat{F}_{s,n}(t) = (1/n)\sum_{i=1}^n \mathbb{I}\{s(X_i) \leq t\}$ denotes the empirical *c.d.f.* of the *r.v.* $s(X)$ and $\widehat{F}_{s,n}^{-1}$ its generalized inverse. In Clémençon and Thomas (2018), for a fixed $s \in \mathscr{S}$, consistency and asymptotic Gaussianity (in sup-norm) of the estimator (8.2.5) has been established, together with the asymptotic validity of a smoothed bootstrap procedure to build confidence regions in the MV space. However, depending on the geometry of the superlevel sets of $s(x)$, it can be far from simple to compute the volumes. In the case where F has compact support, included in $[0,1]^d$ say for simplicity, and from now on it is assumed it is the case, they can be estimated by means of Monte-Carlo simulation. Indeed, if one generates a synthetic *i.i.d.* sample $\{U_1, \ldots, U_m\}$, independent from the X_i's and drawn from the uniform distribution on $[0,1]^d$, which we denote by $\mathscr{U}_d$, a natural estimator of the volume $\widehat{\mathrm{MV}}_s(\alpha)$ is:

$$\widetilde{\mathrm{MV}}_s(\alpha) = \frac{1}{m}\sum_{j=1}^m \mathbb{I}\{s(U_j) \geq \widehat{F}_{s,n}^{-1}(1-\alpha)\} . \tag{8.2.6}$$

Minimization of the empirical area under the MV **curve.** Thanks to the MV curve criterion, it is possible to develop a statistical theory for the anomaly scoring problem. From a statistical learning angle, the goal is to build from training data $X_1, \ldots, X_n$ a scoring function with MV curve as close as possible to MV^*. Whereas the closeness between (continuous) curves can be measured in many ways, the L_1-distance offers crucial advantages. Indeed, we have

$$d_1(s,f) = \int_{\alpha=0}^1 |\mathrm{MV}_s(\alpha) - \mathrm{MV}^*(\alpha)|\,\mathrm{d}\alpha = \int_{\alpha=0}^1 \mathrm{MV}_s(\alpha)\mathrm{d}\alpha - \int_{\alpha=0}^1 \mathrm{MV}^*(\alpha)\mathrm{d}\alpha .$$

Notice that $d_1(s,f)$, $i \in \{1, \infty\}$, is not a distance between the scoring functions s and f but measures the dissimilarity between the preorders they define and that minimizing $d_1(s,f)$ boils down to minimizing the scalar quantity $\int_{\alpha=0}^{1-\varepsilon} \mathrm{MV}_s(\alpha)\mathrm{d}\alpha$, the area under the MV curve. From a practical perspective, one may then learn an anomaly scoring rule by minimizing the empirical quantity

$$\int_0^1 \widetilde{\mathrm{MV}}_s(\alpha)d\alpha .$$

This boils down to maximizing the rank-sum (or Wilcoxon Mann-Whitney) statistic (see Wilcoxon (1945)) given by:

$$\widehat{W}_{n,m}(s) = \sum_{i=1}^n \mathrm{Rank}(s(X_i)) , \tag{8.2.7}$$

where $\mathrm{Rank}(s(X_i))$ is the rank of $s(X_i)$ among the pooled sample $\{s(X_1), \ldots, s(X_n)\} \cup \{s(U_1), \ldots, s(U_m)\}$: $\mathrm{Rank}(s(X_i)) = \sum_{l=1}^n \mathbb{I}\{s(X_l) \leq s(X_i)\} + \sum_{j=1}^m \mathbb{I}\{s(U_j) \leq s(X_i)\}$. Indeed, just like the empirical area under the ROC curve can be related to the rank-sum statistic, we have

$$nm\left(1 - \int_0^1 \widetilde{\mathrm{MV}}_s(\alpha)d\alpha\right) + n(n+1)/2 = \widehat{W}_{n,m}(s) . \tag{8.2.8}$$

In the next section, we introduce more general empirical summaries of the MV curve that are of the form of two-sample rank statistics, just like (8.2.7), and propose to solve the anomaly ranking problem through the maximization of the latter.

8.3 Measuring and optimizing anomaly ranking performance

In this section, a class of anomaly ranking performance criteria are introduced, which can be estimated by two-sample rank statistics. We also emphasize that a natural approach to anomaly ranking consists in maximizing such empirical scalar criteria.

8.3.1 Scalar criteria of performance and two-sample rank statistics

Here we develop the statistical learning framework we propose for anomaly ranking. Let $p \in (0,1)$, we assume that $N \geq 2$ observations are available: $n = \lfloor pN \rfloor$ 'normal' *i.i.d.* observations $X_1, \ldots, X_n$ taking their values in $[0,1]^d$ for simplicity drawn from $F(dx) = f(x)\lambda(dx)$ and $m = N - n$ *i.i.d.* realizations of the uniform distribution $\mathscr{U}_d$, independent from the X_i's. Hence, p represents the 'theoretical' proportion of 'normal' observations among the pooled sample. Let a class of scoring functions $\mathscr{S}_0 \subset \mathscr{S}$ such that, for all $s(x)$, we consider the mixture distribution $G_s = pF_s + (1-p)\lambda_s$ and its empirical counterpart $\widehat{G}_{s,N}(t) = (1/n)\sum_{i=1}^{n} \mathbb{I}\{s(X_i) \leq t\} + (1/m)\sum_{j=1}^{m} \mathbb{I}\{s(U_i) \leq t\}$. Notice that since $n/N \to p$ as N tends to infinity, the quantity above is a natural estimator of the *c.d.f.* G_s. We refer to the *scored* random samples for $\{s(X_1), \ldots, s(X_n)\}$ and $\{s(U_1), \ldots, s(U_m)\}$. Therefore, motivated by Eq. (8.2.8), Definition 80 below provides the class of *W_ϕ-performance criteria* we consider in the subsequent procedure.

Definition 80. *Let $\phi : [0,1] \to \mathbb{R}$ be a nondecreasing function. The 'W_ϕ-ranking performance criterion' with 'score-generating function' $\phi(u)$ based on the mixture cdf $G_s(dt)$ is given by:*

$$W_\phi(s) = \mathbb{E}[(\phi \circ G_s)(s(X))] \,. \tag{8.3.1}$$

One can naturally relate this generalized form to the MV curve, justifying this choice of scalar performance criteria as summaries of the MV curve, through the equality

$$W_\phi(s) = \int_0^1 \phi\left(1 - p\alpha - (1-p)\mathrm{MV}_s(\alpha)\right) d\alpha \,. \tag{8.3.2}$$

Equipped with the two random samples, the following Definition 81 provides an empirical counterpart, that generalizes the empirical summaries of the MV curve *via* collections of two-sample linear rank statistics. Precisely, for a given mapping $s(x)$, we allow to weight the sequence of 'normal ranks' *i.e.* the ranks of the scored 'normal' instances among the pooled sample, by means of a *score-generating* function.

Definition 81. (TWO-SAMPLE LINEAR RANK STATISTICS) *Let $\phi : [0,1] \to \mathbb{R}$ be a nondecreasing function. The two-sample linear rank statistics with 'score-generating function' $\phi(u)$ based on the random samples $\{X_1, \ldots, X_n\}$ and $\{U_1, \ldots, U_m\}$ is given by:*

$$\widehat{W}_{n,m}^{\phi}(s) = \sum_{i=1}^{n} \phi\left(\frac{\mathrm{Rank}(s(X_i))}{N+1}\right) \,, \tag{8.3.3}$$

where $\mathrm{Rank}(t) = N\widehat{G}_{s,N}(t) = \sum_{i=1}^{n} \mathbb{I}\{s(X_i) \leq t\} + \sum_{j=1}^{m} \mathbb{I}\{s(U_j) \leq t\}$.

Optimality. Briefly, we refer to the comprehensive analysis of the general class of criteria in Clémençon et al. (2021), that establishes the theoretical guarantees for the consistency of the two-stage procedure we detail in the following subsection. Importantly, the set of optimal maximizers of the empirical W_ϕ-criteria coincides with the nondecreasing transforms of the likelihood ratio, just like for the MV curves, as shown thourgh the Eq. (8.3.2). The optimal set $\mathscr{S}^*$ derived in Eq. (8.2.1) underlines the implicit characterization that inherits an outlier: the lower the scalar score is and the likelier anomalous the observation can be considered. Also, the notion of distance induced by the rank-based criteria is in fact directly related to the distribution of the 'normal' sample compared to the Uniform one.

Choosing ϕ. As foreshadowed above, the choice of the score-generating function is an asset of this class of criteria as it provides a flexibility *w.r.t.* the weighting of the area under the MV curve. Indeed, its minimization directly implies the maximization of the W_ϕ-criterion (see Eq. (8.3.2)), recalling the nondecreasing variation of $\phi(u)$. Therefore, one can hope to recover at best the MV* curve by the right choice of $\phi(u)$, especially when the initial sample is noisy. Additionally, when going back to the problem of learning to rank the (possible abnormal) instances, it is an advantage to weight the ranks accordingly.

First, we recall the simplest uniform weighting of each 'normal' rank with $\phi(u) = u$. It parenthetically yields to Eq. (8.2.8), of continuous version: $W(s) = p/2 + (1-p)(1 - \int_0^1 \mathrm{MV}_s(\alpha)\mathrm{d}\alpha)$, where the area under the MV curve is clearly computed. Other functions were introduced in the literature related to classic univariate two-sample rank statistics. Figure 5.1 gathers classical nondecreasing score-generating functions broadly used for two-sample statistical tests (refer to Hájek (1962)).

8.3.2 The two-stage procedure

In this paragraph, we detail the two-stage procedure, where we assume that both the framework and assumptions detailed in the previous subsection are adopted. We define the test sample as the set of *i.i.d.* random variables $\{X_1^t, \ldots, X_{n_t}^t\}$, with $n_t \in \mathbb{N}^*$, *a priori* drawn from $F(dx)$. The goal pursued is to distinguish among the test sample, the instances the most likelier to be anomalous. In particular, we propose a first step $(1.)$ that outputs an optimal ranking rule $\hat{s}_{n,m}(x)$, in the sense of the maximization of the rank statistics of Eq. (81). Then, in the second step $(2.)$ and equipped with this rule, the instances of the test sample are optimally ranked by increasing order of similarity *w.r.t.* the X's. We also choose to watch a number of $n_{lowest} \in \mathbb{N}^*$ worst ranked instances *i.e.* of lowest empirical score. The procedure is detailed in the following Fig. 8.2. By means of the recalled theoretical guarantees proved in Clémençon et al. (2021), it results to the asymptotic consistency of step $(1.)$ as well as its nonasymptotic consistency with high probability, under some technical assumptions.

8.4 Numerical experiments

In this section, we illustrate the procedure promoted along the chapter through numerical experiments on imbalanced synthetic data. As these experiments are mainly here to support our methodology, we propose for the step $(1.)$ to learn the empirical maximizer $\hat{s}_{n,m}$ by means of a regularized classification algorithm. At a technical level, we would ideally like to replace usual loss criterion such as the BCE (Binary Cross-Entropy) loss by our tailored objective W_ϕ. Unfortunately, the latter is not smooth and of highly correlated terms, which results in many challenges regarding its optimization. In order to incorporate W_ϕ and still keeping good performances, we *(i)* use a regularized proxy of it and *(ii)* incorporate the regularized criterion in a penalization term. The second point allows to drive the

ANOMALY RANKING PROCEDURE

Consider the 'normal' *i.i.d.* random sample $\{X_1, \ldots, X_n\}$ of unknown probability probability function $F(dt)$ defined on a feature space included $[0,1]^d$. Consider a new random sample $\{X_1^t, \ldots, X_{n_t}^t\}$ *a priori* drawn from $F(dt)$, $n_t \in \mathbb{N}^*$, independent of the X's. Let $\mathscr{S}_0 \subset \mathscr{S}$ a class of scoring functions and ϕ a score-generating function. Set $n_{lowest} \in \mathbb{N}^*$.

1. Maximizing the W_ϕ-ranking performance criterion.

(i) Generate a *i.i.d.* random sample $\{U_1, \ldots, U_m\}$ from $\mathscr{U}_d(dt)$, independent of $\{X_1, \ldots, X_n\}$,

(ii) Output the optimal empirical scoring rule $\hat{s}_{n,m} = \arg\max_{s \in \mathscr{S}_0} \widehat{W}_{n,m}^{\phi}(s)$ based on the two samples $\{X_1, \ldots, X_n\}$ and $\{U_1, \ldots, U_m\}$.

2. Ranking anomalies.

(i) Compute the empirical scores of the test sample $\{\hat{s}_{n,m}(X_1^t), \ldots, \hat{s}_{n,m}(X_{n_t}^t)\}$,

(ii) Define as anomalous the n_{lowest} observations of lowest empirical scores among the sequence $\hat{s}_{n,m}(X_i^t)$, $i \leq n_t$.

Result. Output the set of anomalous observations with their corresponding rank.

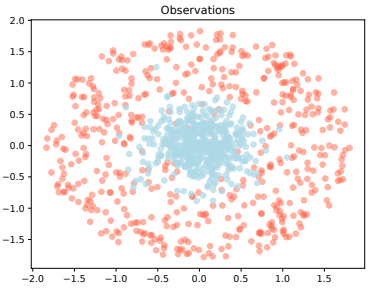
Figure 8.2. Two-stage procedure for learning to rank anomalies.

learning with a usual BCE loss, which asymptotically amounts to estimate the conditional probability $\mathbb{P}(y = 1 \,|\, X)$, while considering W_ϕ.

Data generating process. We generated the 'positive' sample by *i.i.d.* Gaussian variables $X_1, \ldots, X_n$, $n = 1000$, in dimension $d = 2$, centered and with covariance matrix $0.1 \times I_2$ (where I_2 is the identity matrix). We chose the Gaussian law for its attractive structure and in particular for its symmetry, it can be a reasonable choice in many situations where the data at hand are indeed well structured. We then sampled the 'negative' sequence of *i.i.d. r.v.* $U_1', \ldots, U_m'$, $m = 500$, from the following radial law, expressed in terms of its density in polar coordinates:

$$\text{RadLaw}_{\alpha,\beta} : (v,r) \in \mathbb{S}^{d-1} \times (0,1) \mapsto \frac{1}{\text{Area}(\mathbb{S}^{d-1})} dv \times \frac{1}{B(\alpha,\beta)} r^{\alpha-1}(1-r)^{\beta-1} dr \,,$$

where $\alpha, \beta > 0$ are two tunable parameters, $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d,\ \|x\| = 1\}$ is the unit sphere, and where $B(\alpha,\beta) = \int_0^1 r^{\alpha-1}(1-r)^{\beta-1} dr$. In other words, v is uniformly sampled in the unit sphere and r has Beta law with parameters α and β. Notice that $\alpha = \beta = 1$ corresponds to the Uniform law and that, when $\beta = 1$, the law puts more mass around 1 as $\alpha > 1$ increases. In our experiment, we choose $\alpha = 3$ and $\beta = 1$. Denoting by $\text{rad} = \max_{1 \leq i \leq n} ||X_i||$, we finally obtained m 'synthetic outliers' $U_1, \ldots, U_m$ defined by $U_i = (\text{rad} + \varepsilon) \times U_i'$, with $\varepsilon = 0.01$. To simplify the notations, we denote by $\mathbf{Z}_{train}$ the concatenation of the X_i's and the U_i's. We also denote by $\mathbf{y}_{train}$ the labels, where we choose to assign the label 1 (*resp.* 0) to the 'positive' (*resp.* 'negative') sample. Figure 8.3 illustrates both data generating processes. For the test set, we generated similarly a sequence of $n_t = 400$ *i.i.d.* Gaussian *r.v.* $X_1^t, \ldots, X_{n_t}^t$ from the same Gaussian law as the 'positive' sample, and a *i.i.d.* random sequence $U_1^t, \ldots, U_{m_t}^t$, $m_t = 100$, drawn from the law $\text{RadLaw}_{\alpha_t,\beta_t}$, with $\alpha_t = 2.$ and $\beta_t = 1.$, dilated by a factor $(\text{rad} + \varepsilon)$.



(a) Train data. $(n,\ m) = (1000,\ 500)$.

(b) Test data. $(n_t,\ m_t) = (400,\ 100)$.

Figure 8.3. Data visualization for the two generating processes. The Gaussian observations are represented in blue. The 'synthetic outliers' samples drawn from the radial law are represented in red. The left figure (a) corresponds to the train dataset, the right (b) to the test dataset.

Metrics. Once the algorithm that learns a (renormalized) optimal scoring function $\hat{s}_{n,m} : \mathbb{R}^d \to (0,1)$ has been trained (*i.e.* step $(1.)$), we score the test data with $\hat{s}_{n,m}$ and compute the proportion of true outliers among the n_{lowest} points having lowest scores (*i.e.* step $(2.)$). We let n_{lowest} varies in $\{25,50,75,100\}$. Formally, if $\xi_1 \preccurlyeq \cdots \preccurlyeq \xi_{n_t+m_t}$ denote the points X_i^t and U_i^t sorted by scores, *i.e.* the ordered sequence based on $\hat{s}_{n,m}(\mathbf{Z}_{test,1}), \ldots, \hat{s}_{n,m}(\mathbf{Z}_{test,n_t+m_t})$, we compute the following accuracy

$$\mathrm{Acc}_{n_{lowest}} = \frac{1}{n_{lowest}} \sum_{i=1}^{n_{lowest}} \mathbb{I}\{\xi_i \in \{U_1^t, \ldots, U_{m_t}^t\}\} \ . \tag{8.4.1}$$

Neural Network. We trained a neural network MLP composed of one hidden layer of size $2 \times d$, a ReLu activation function and whose last layer is a Sigmoid function, computing the desired score. For each $n_{epoch} = 30$ epochs, we use the following training scheme:

1. Each sample of $(\mathbf{Z}_{train},\ \mathbf{y}_{train})$ is individually passed through the network, the BCE loss is computed[2] and a backpropagation step is performed,

2. At the end of each epoch, the whole batch of the training dataset $(\mathbf{Z}_{train},\ \mathbf{y}_{train})$ is passed through the network and we computed the Binary Cross Entropy loss, denoted by BCE, and the following proxy of W_ϕ

$$\widehat{W}_{n,m}^{\phi} = \sum_{i=1}^{n} \phi\left(\frac{(n+m) \times \mathrm{MLP}(X_i) + 1}{n+m+1}\right) \ .$$

   In our experiments, we choose $\phi(u) = u$ and $\phi_{u_0}(u) = u\mathbb{I}\{u \geq u_0\}$ with $u_0 = 0.7$, as defined in section 8.3.1. We then compute the regularized loss $\mathrm{BCE} - \lambda \widehat{W}_{n,m}^{\phi}$, where λ is a hyperparameter in $\{0, 0.01, 0.1, 1, 10\}$.

The training procedure of the Neural Net is summarized in the Algorithm 5.

Repetitions. We repeat $B = 100$ times the procedure, each time computing the accuracy metric defined above.

[2] Remember it is given by $-y \ln \hat{y} - (1-y)\ln(1-\hat{y})$, where $\hat{y} = \mathrm{MLP}(X)$.

Algorithm 5: Training of the Neural Network

Data: $(\mathbf{Z}_{train}, \mathbf{y}_{train})$.
Input: Network MLP, number of epochs n_{epoch}, penalization strength λ.
Result: Trained network.

```
for n = 0, ..., n_epoch do
    for X, y ∈ Z_train, y_train do
        compute ŷ = MLP ;
        compute BCE = BCE(ŷ, y), backpropagate and zero_grad ;
    end
    compute ŷ = MLP(Z_train) ;
    compute BCE = BCE(ŷ, y) and Ŵ^φ_{n,m} ;
    compute the regularized loss BCE − λŴ^φ_{n,m}, backpropagate and zero_grad ;
end
```

Visualization and results. In this section, we only display the results obtained with $\phi(u) = u$ since they are very similar to the one obtained with $\phi(u) = u\mathbb{I}\{u \geq u_0\}$. This is probably due to the very simple framework adopted for the data generating process and further investigations would be of interest.

For the first learning loop, we saved the evolution of the BCE losses, for all values of λ, computed at each epoch together with the W_ϕ proxy and the accuracy metric for $n_{lowest} = 75$. As displayed in Figure 8.4, one can see that the incorporation of the empirical W_ϕ criterion in the penalization term improves the performances for a well chosen parameter λ. For instance, $\lambda \in \{1, 10\}$ output the best results in this setting.

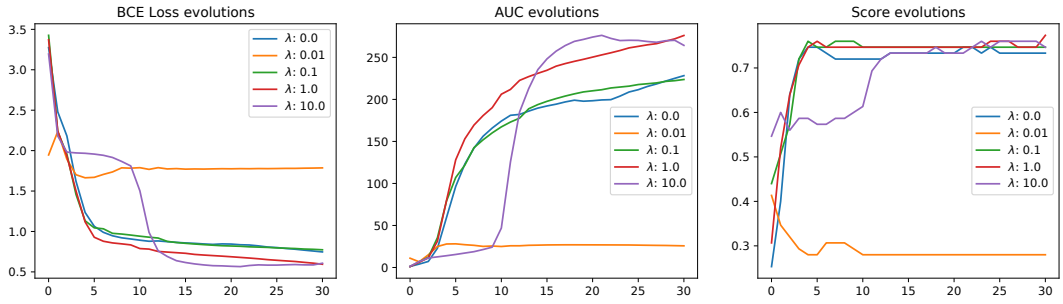


Figure 8.4. Evolutions of the BCE loss, the AUC proxy and the accuracy for $n_{lowest} = 75$ in function of the epochs, for $\phi(u) = u$ and all values of the hyperparameter $\lambda \in \{0, 0.01, 0.1, 1, 10\}$.

At the end of the training, we select the network having the highest empirical W_ϕ score, which here corresponds to choosing $\lambda = 1$. We then score the initial observations $X_1, \ldots, X_n$ and display in Figure 8.5 the points with an intensity varying from red to blue as the score increases from 0 to 1. The fact that the red points are on the sides of the dataset empirically validates our methodology. We represent in Fig. 8.6 the averaged mass volume curve together with standard deviation computed for $\lambda = 1$ over $B = 50$ repetitions. Table 8.1 gathers the results averaged over $B = 50$ repetitions. Notice that these results support the soundness of our approach. Indeed, the area under the MV curve is minimized and the proportion of detected outliers is high even when n_{lowest} increases.

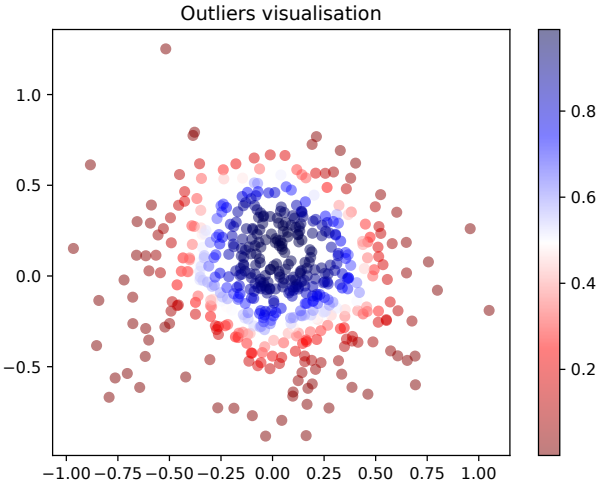
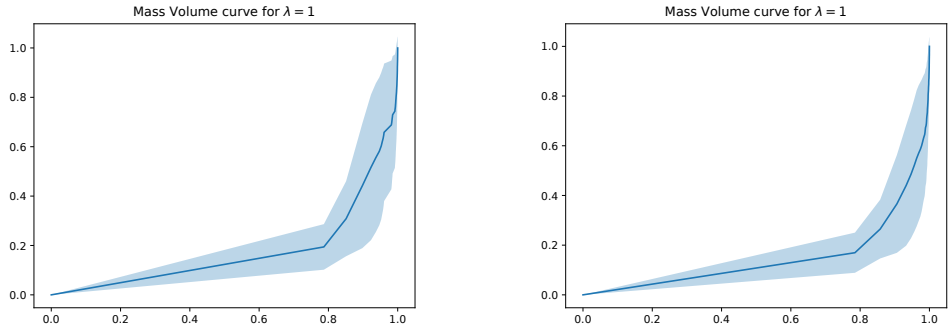


Figure 8.5. A heatmap of the scores for $\phi(u) = u$.

n_{lowest}	25	50	75	100
$\text{Acc}_{n_{lowest}}$	0.91 ± 0.13	0.84 ± 0.15	0.74 ± 0.15	0.64 ± 0.13

Table 8.1. Tabular view of the empirical accuracy $\pm$ its standard deviation, when n_{lowest} varies in $\{25, 50, 75, 100\}$, with $\lambda = 1$.



(a) $\lambda = 1$ and $\phi(u) = u$.

(b) $\lambda = 1$ and $\phi(u) = u\mathbb{I}\{u \geq u_0\}$.

Figure 8.6. Empirical Mass-Volume curves.

8.5 Conclusion

In this chapter, we promoted a binary classification approach to the problem of learning to rank anomalies. We established a clear theoretical link between these two machine learning tasks through the study of the mass-volume curve. In particular, our procedure is robust with respect to imbalanced datasets through the choice of the parameter p that is chosen initially in practice. Previous results obtained in Chapter 5 support the effectiveness of our methodology. Moreover, we illustrate our method with numerical experiments of synthetic data.

9 | The Two-sample Problem Applied to Biomedical Studies

Abstract. Falling in Parkinsonian syndromes (PS) is associated with postural instability and consists a common cause of disability among PS patients. Current posturographic practices record the body's center-of-pressure displacement (statokinesigram) while the patient stands on a force platform. Statokinesigrams, after appropriate processing, can offer numerous posturographic features. This fact, although beneficial, challenges the efforts for valid statistics via standard univariate approaches. In this work, 123 PS patients were classified into fallers (PS_F) or non-faller (PS_{NF}) based on the clinical assessment, and underwent simple Romberg Test (eyes open/eyes closed). We developed a non-parametric multivariate two-sample test (ts-AUC) based on machine learning, in order to examine statokinesigrams' differences between PS_F and PS_{NF}. We analyzed posturographic features using both multiple testing with p-value adjustment and ts-AUC. While ts-AUC showed significant difference between groups (p-value $= 0.01$), multiple testing did not agree with this result (eyes open). PS_F showed significantly increased antero-posterior movements as well as increased posturographic area compared to PS_{NF}. Our study highlights the superiority of ts-AUC compared to standard statistical tools in distinguishing PS_F and PS_{NF} in multidimensional space. Machine learning-based statistical tests can be seen as a natural extension of classical statistics and should be considered, especially when dealing with multifactorial assessments.

Contents

9.1 Introduction

Postural control is the capacity of an individual to maintain a controlled upright position. Falls have been reported as one of the major causes of injury among elderly and more importantly among patients of balance-related disorders, such as Parkinsonian syndromes (PS). It has been estimated that one third of the population over 65 years-old faces minimum one fall per year Tinetti (2003a). Falls promote the decrease in mobility, problems of autonomy in daily activities (bathing, cooking, etc.), or even death Melzer et al. (2004); Tinetti (2003a). Taking also into consideration the aging of many modern societies, accurate risk assessment has become a major challenge with huge socio-economic impact Stevens et al. (2006).

Force platforms are one of available acquisition tools of clinical researchers for the assessment of postural control. Such platforms record the displacement of the center of pressure (CoP) applied by the whole body in time while the individual stands upon it and follows the clinician's instructions. These CoP trajectories, usually called statokinesigrams, have been widely used in assessing the balance disorder in healthy or PS populations. It has been shown that CoP displacement characteristics can reflect individuals' postural impairment when special acquisition protocols are followed Chagdes et al. (2009); Mancini et al. (2012a); Melzer et al. (2004).

Clinical research often aims to find the significant differences between fall-prone individuals and others who have not yet manifested important balance impairment. Researchers usually compute several features using signal processing techniques and evaluate their usefulness relying on a variety of available univariate tests, such as the Student's t-test, Kolmogorov–Smirnov or Mann-Whitney Wilcoxon. However, usually in experimental works, where pre-planned hypotheses are not well-fixed, multiple univariate tests are applied consecutively in order to find the features that separate significantly the two groups. The aforementioned multiple testing scheme has been part of a well-known scientific debate Feise (2002), mainly criticized for the increased probability of reporting a false-positive finding. More specifically, it has been reported that for alpha level $\alpha = 0.05$, it is possible that 1 in 20 relationships may be statistically significant but not clinically meaningful Feise (2002). Thus, several biostatisticians recommend to disclose all the elements of the conducted analysis, and not only the elements that found to be significant. The violation of this recommendation and the regular misuse of those tests Thiese et al. (2015) combined with the relatively small available

cohorts, may lead to false conclusions and as a consequence to a significant lack of clinical consensus or at least delay in reaching it. Well-known adjustments have been proposed in order to limit the aforementioned probability of a false-positive finding (such as Bonferroni corrections) but they have been reported as conservative compromises (due to the significant increase of the probability for false-negative output) Feise (2002) that do not constitute a satisfactory solution Perneger (1998a). Other corrections (more powerful than Bonferroni) such as Hommel (1988), Hochberg (1988) and Holm (1979) (in descending power order Gou et al. (2014)) have been also proposed.

Classic statistical tests are very sensitive on the size of the available dataset. The generalization of any result is not safe when only relatively small populations are available (see Wood et al. (2014) for the high risk of making false conclusions). In order to reduce this sensitivity, machine learning algorithms assess their results using cross-validation schemes. Briefly, an algorithm trains a model that 'learns' to solve the problem in a randomly selected part of the dataset (called training-set), and then tests whether it can be effective on the rest of the 'unseen' data (test-set). The learning and validation process is repeated multiple times and performance metrics are averaged. In the context of multidimensional datasets with binary labels $\{-1,+1\}$, the idea of assessing the separability of two groups is based on the aforementioned learning and validation scheme. The learning process sets the criteria in order to rank the population in the test-set by means of a scoring function s. Those who are ranked at the top of the list will be considered to belong to the positive class Clémençon et al. (2009). The machine learning community has recently made significant progress in this topic Bach et al. (2008); Chen and Qin (2010); Clémençon et al. (2005); Gretton et al. (2012b), especially related to the design of appropriate criteria for the characterization of the ranking performance and/or meaningful extensions of the Empirical Risk Minimization (ERM) approach to this framework Agarwal et al. (2005); Cortes and Mohri (2004). In a large part of these efforts, the well-known criterion of the area under the ROC curve (AUC) is considered as the gold standard for measuring the capacity of a scoring function to discriminate groups of populations Clémençon et al. (2009). Briefly, in the setting of two-sample statistical testing, an algorithm 'learns' the rule that maximizes the AUC between the two groups in the training-set, and then tests the applicability of this rule to the test-set during the validation process.

Unfortunately, to the best of our knowledge, these novel advancements in statistical testing remain largely unexploited by the parkinsonism-related community. The lack of common language and proper methodological simplifications to make the approaches easy to understand by clinical researchers are possibly the major reasons for such an observed distance.

In postural research, simple acquisition protocols (such as the basic Romberg test) have been reported to contain inconclusive information to evaluate sufficiently the postural control of an individual Palmieri et al. (2002). However, only recently, works proposed that a combination of multiple global features, derived from CoP trajectories using data mining techniques, might be advantageous in order to classify fallers and non-fallers. Earlier works Audiffren et al. (2016); Bargiotas et al. (2018), showed that although none of the features alone could classify effectively elderly fallers/non-fallers (i.e. weak classifiers), yet combining all features through non-linear multi-dimensional classification gave significant results. It is suggested that the shape of the decision surface lies indeed in a multidimensional space and should be learned using multiple features at once. As a consequence, the above findings raise reasonable questions about the ability of traditional statistical tools and testing protocols to fully reveal and exploit the existing associations.

The objective of the present study is to propose an easy-to-use and -interpret two-sample hypothesis testing approach, in an attempt to address some the aforementioned difficulties of clinical research. Our contribution is to propose a new variation of a multivariate two-sample test through AUC maximization, which was originally theoretically established in Clémençon et al. (2009), and test it to a PS population which includes two groups: fallers (PS_F) and non-fallers (PS_{NF}). We intend

to highlight the benefits that one might have by using such kind of two-sample analysis in the presence of multiple features, and demonstrate the contradicting conclusions that a traditional statistical analysis (hypothetical future clinical study) might have had compared to the proposed method. In addition, we performed comparative performance in simulated synthetic data in order to strengthen the evidence that the proposed approach is statistically sound and consistent. Therefore, we decided to conduct such a study, providing it though in the Section 9.5.1 in order to keep the main text focused on the problem-specific results in which we are primarily interested.

9.2 Materials and methods

9.2.1 Balance measurements and fall assessment

Our dataset comes from the Neurology department of the HIA, Percy hospital (Clamart, France), and includes 123 patients (78.7 ± 5.4 years-old, Table 9.1) who suffered from Parkinsonian syndromes. PS patients that suffered from other comorbidities (such as vestibular and proprioceptive impairements) were not included in the study. Following the acquisition protocol, patients were asked to remove their shoes and to maintain upright position on a force platform keeping their eyes open and their arms at the side. The CoP trajectory was recorded for 25 seconds at that stance. After that, patients were asked to close their eyes maintaining their upright position. After a ten-second pause, clinical experts recorded 25 additional seconds with eyes closed (Figure 9.1).

Characteristics	**Non-Fallers**	**Fallers**
Population size	99	24
Age	78.8 ± 5.3	78.5 ± 5.9
Gender	M:71/W:28	M:16/W:8
UPDRS III total score	23.6 ± 11.9	26.3 ± 11.1
Disease duration	4.7 ± 3.5	5.7 ± 4.2

Table 9.1. Population characteristics: the 123 patients included in the study.

Statokinesigrams were acquired using a Wii Balance Board (WBB) (Nintendo, Kyoto, Japan), which has been found to be a suitable and convenient tool for the clinical setting Clark et al. (2010); Leach et al. (2014), and the newly proposed portable package developed in our laboratory. Statokinesigram from WBB are sent to the clinician's professional Android tablet via Bluetooth connection. Acquired signals are sent (after anonymization and encryption) to a central database for high level processing (computation of features associated to postural control and application of appropriate algorithms Audiffren et al. (2016); Bargiotas et al. (2018, 2019)), and the demanded results are communicated to the clinician online. Since the WBB records the CoP trajectories at non-stable time resolution, the acquired statokinesigrams are resampled at 25Hz using the SWARII algorithm Audiffren and Contal (2016).

In order to label the participants, a questionnaire (implemented to the Android tablet) was filled for every subject registering information about falls during the last six months prior to the examination. As in previous works Zecevic et al., participants were labeled as fallers (PS_F) if they had come to a lower level near the ground unintentionally at least once during that period. Twenty-four (24) patients were labeled as fallers. Any useful information about the conditions of falls were registered. The clinical trial registered at ANSM (ID RCB 2014-A00222-45) was approved by the following ethics committee/institutional review board(s): 1) Ethical Research Committees (CPP), Ile-de-France, Paris VI; 2) French National Agency for the Safety of Medicines and Health Products (ANSM); 3) National Commission on Informatics and Liberty (study complies with the MR-001). All research

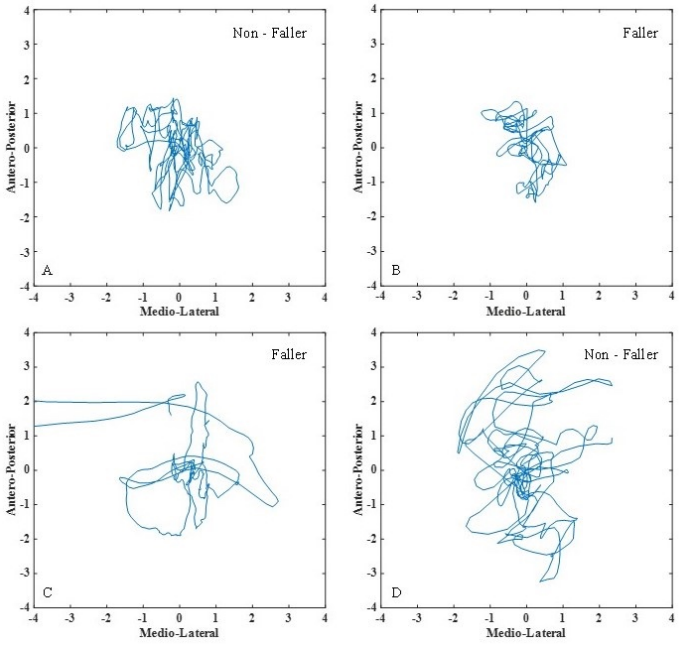


Figure 9.1. Examples of statokinesigrams from fallers and non-fallers. The x-axis is the medio-lateral (ML) movement and the y-axis is the antero-posterior (AP) movement of the body in centimeters (cm) during the acquisition. As it can be observed, fallers and non-fallers are not easily distinguishable by examining visually their statokinesigrams.

was performed in accordance with relevant guidelines and regulations. After information and allowing adequate time for consideration, written informed consent was obtained from all participants before being included in the study.

9.2.2 Choice of posturographic features

Our analysis included only features that were computed on the two-dimensional CoP displacement and have been previously proposed as indicators of postural impairment Błaszczyk et al. (2007); Melzer et al. (2004); Muir et al. (2013). Table 9.2 provides the names, measuring units, and descriptions (where needed) for the features that were included in the test.

9.2.3 Two-sample test through AUC optimization (ts-AUC)

We applied a bootstrap aggregation classification, in particular a random forest (RF) Breiman (2001) that comprises several decision trees (DTs). Therefore, in the development of each DT, only a part of the whole dataset does participate (in-bag) while the other part is left out (out-of-bag, or OOB). Consequently, the OOB subset can be used as test-set for the particular DT. In our approach, instead of the originally proposed testing method based on data splitting, we used the predictions of the OOB population Breiman (1996). The number of DTs was large enough ($T = 200$) compared to the actual population. The individuals can be selected in different OOB sets more than once. Every time an individual is part of an OOB set, the corresponding DT outputs the probability for him/her being a PS_F or a PS_{NF}. This is computed as the fraction of individuals of the positive class (fallers) in the tree leaf where he/she reaches. Thus, his/her final score is given by the average of the posterior probabilities over the trees he/she was part of the OOB set (see Figure 9.2). Averaged posterior probabilities (P) of the positive class (fallers) are used in order to compute the Mann-Whitney U-test

Feature	Unit	Description
`RangeX`	cm	–
`MaxX`	cm	Maximum medio-lateral displacement (right)
`MinX`	cm	Minimum medio-lateral displacement (left)
`VarianceX`	cm^2	–
`VelocityX`	cm/s	Average instant x-axis velocity of CoP changes
`AccelerationX`	cm/s^2	Average instant x-axis acceleration of CoP changes
`F95X`	Hz	Frequency below which 95% of the x-axis CoP trajectory's energy lies
`RangeY`	cm	–
`MaxY`	cm	Maximum antero-posterior displacement (front)
`MinY`	cm	Minimum antero-posterior displacement (back)
`VarianceY`	cm^2	–
`VelocityY`	cm/s	Average instant y-axis velocity of CoP changes
`AccelerationY`	cm/s^2	Average instant y-axis acceleration of CoP changes
`F95Y`	Hz	Frequency below which 95% of the y-axis CoP trajectory's energy lays)
`DistC`	cm	Instant distance from the center of the trajectory
`EllArea`	cm^2	Confidence ellipse area that covers the 95% of the trajectory's points
`AngularDeviation`	degrees	Average of the angle of deviation

Table 9.2. Computed features derived from the CoP displacement during the acquisitions.

statistic, denoted by U as proposed in the theoretical work of Clémençon et al. (2009) . The empirical AUC for the chosen hyperparameters is given by $U/(N_{\mathrm{F}} \cdot N_{\mathrm{NF}})$. Briefly, the null hypothesis, H_0, and the alternative one, H_1, are expressed as follows:

$$\text{“}\mathrm{H}_0 : \mathrm{AUC}^* = \frac{1}{2}\text{”} \quad \textit{vs.} \quad \text{“}\mathrm{H}_1 : \mathrm{AUC}^* > \frac{1}{2}\text{”} . \tag{9.2.1}$$

The OOB percentage was fixed to 36.8% of the included population. Searching the empirical AUC^* (maximal AUC), the hyperparameters that are optimized are the leaf-size LS and the number of features to be used to build each tree M. We avoided a greedy approach using a Bayesian optimization process, where only relatively shallow ($7 < LS < 20$) and simple ($M < 9$) DTs were allowed to be tested. The averaged posterior probabilities of the *star model*, where $\mathrm{AUC} = \mathrm{AUC}^*$, are used to compute the scoring function (and the p-value) through a univariate Mann-Whitney Wilcoxon (MWW from now on) test on the whole available dataset (see Algorithm 6 and Figure 9.2).

9.2.4 Out-of-bag feature importance

Additionally, the proposed algorithmic modifications allow the assessment of the importance of each feature to the ts-AUC's final decision. We estimated the out-of-bag feature importance by permutation. Briefly, the more important a feature is, the higher its influence (i.e. the increase) would be to the model's error after feature's random permutation at the OOB subset. The permutation of a non-influential feature will have minimum, or no effect at all, on the model's error. Having D features in the dataset and T trees in the RF model, the influence of feature $j \in \{1, ..., D\}$ is computed as:

$$I_j = \frac{d_j}{\sigma_j} , \tag{9.2.2}$$

where d_j is the average change of model error after the permutation of feature j, and σ_j is the standard deviation of the above change. Important to explain that every feature j participates only to the training of a subset of the trees of the RF. Therefore, d_j and σ_j are derived by those trees in which the feature j was selected to participate in their training.

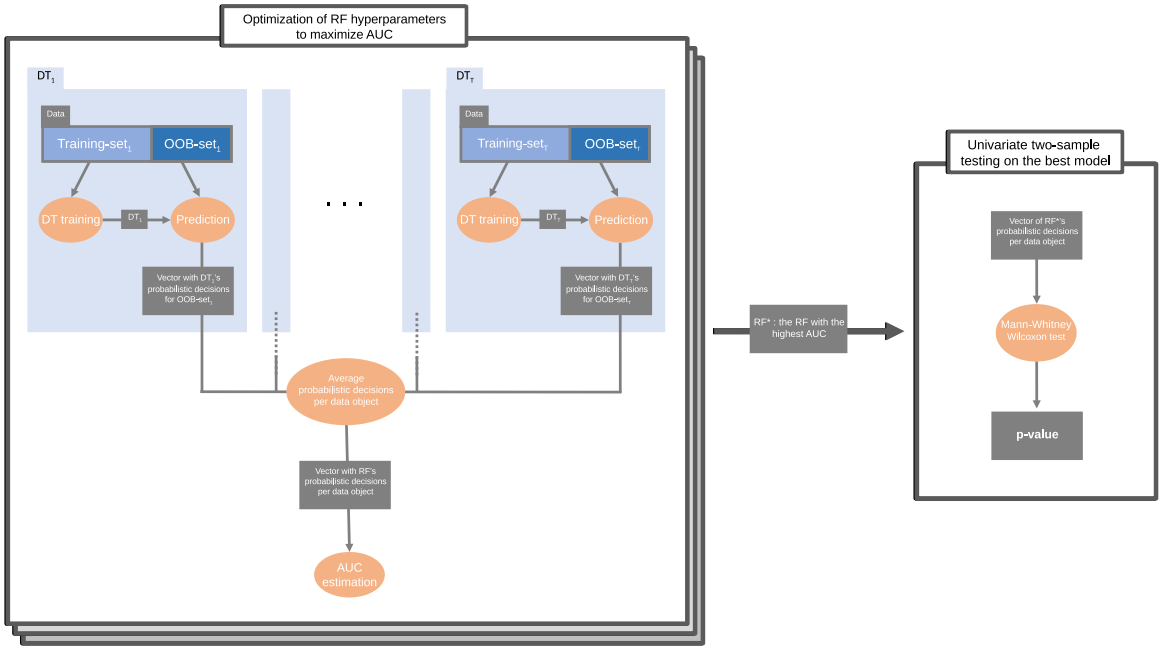


Figure 9.2. Scheme of the ts-AUC algorithm. In order to find the AUC* (maximal AUC), a number of Random Forests (RFs). For the RF* with the best AUC*, the univariate Mann-Whitney Wilcoxon non-parametric two-sample test is applied on the average posterior probability values of the whole population.

Algorithm 6: The proposed ts-AUC statistical test.

Data: X and Y are the points' coordinates of the trajectory (statokinesigram).
Result: AUC^*, RF^*, P^*, $p\text{-}value^*$.

Step 1: Exploration of the space of hyperparameters.

for $i \in LS$ **do**
  for $j \in M$ **do**
    $RF = \text{RandForest}(X, Y, LS_i, M_j)$;
    $P = \text{OOBpredict}(RF_{i,j})$;
    $U = \text{Mann\_Whitney\_Utest\_Statistic}(P)$;
    $AUC_{i,j} = \text{AUCestimation}(U, Y)$;
  end
end

Step 2: Choose the best model and apply MWW.

Set $(i^*, j^*) = \arg\max_{i \in LS, j \in M} AUC_{i,j}$;
Set $AUC^* = AUC_{i^*,j^*}$;
Set $RF^* = \text{RandForest}(X, Y, LS_{i^*}, M_{j^*})$;
Set $P^* = \text{OOBpredict}(RF^*)$;
$p\text{-}value^* = \text{MWW}(P^*, Y)$.

Since our objective is to enhance interpretability of results, our feature importance analysis aims to identify all the important features, even those which are redundant or colinear, rather than finding a parsimonious set of important features. Hence, we followed the additional procedure proposed in

Genuer et al. (2010) especially for interpretation purposes. Briefly, we computed the AUC of the OOB (AUC_{OOB}) of RFs starting from the most important feature, and adding progressively all the others in descending importance order. The best model is the smallest model (less features) with an AUC_{OOB} higher than the maximum AUC_{OOB} reduced by its empirical standard deviation (based on 20 runs).

9.2.5 Experimental settings

We compare the results obtained by the proposed ts-AUC with the Maximum Mean Discrepancy test (MMD-test) Gretton et al. (2012a), which is a well-established multivariate test and state-of-the-art in terms of performance. The MMD measures the maximum difference between the mean of two data samples, in the space of probability measures of a Reproducing Kernel Hilbert Space (RKHS). Practically, this test uses the unbiased squared MMD statistic. It has been proven to be highly efficient and easy to use (a package with kernel optimization is provided in Sutherland et al. (2017)).

In addition, we compare the results of ts-AUC with standard statistical testing approaches which are usually used in clinical studies. We checked the p-values of all 17 features (i.e. $D = 17$) with the labels {'faller'/'non-faller'} using the non-parametric Mann-Whitney Wilcoxon test. Typically, clinicians would report those features which were found statistically significant (e.g. with p-value $< \alpha = 0.05$) and any interesting non-significant finding.

In order to prevent the increase of the false positive probability due to the large number of tested hypotheses, p-value adjustment procedures are applied. We use the Bonferroni correction, which is the most widely used p-value adjustment in biomedical research. Moreover, after taking into account the criticism that Bonferroni has received Perneger (1998a), we also apply alternative approaches such as Hommel (1988), Hochberg (1988), Holm (1979) and Bonferroni corrections.

We assess the effect of population size to the final result by performing the following two additional experiments:

1. We progressively decrease, uniformly at random, the population size by a step of 10% (from 95% to 35%).

2. We progressively reduce, uniformly at random, the number of PS_{NF} by a step of 10% (from 95% to 35%).

At every step, the analysis of each case runs 12 times and the percentages of significant results were compared (see Figure 9.6 and Figure 9.5).

Finally, to enhance further our conclusions, we compared the behaviour of the tests to simulated groups with various populations (N from 100 to 200), various levels of separation (difference in mean values) and various class proportions between the two groups (50/50, 70/30, 90/10, percentages of positives/negatives). These results can be found in the Appendix (see Figures 9.7, 9.8, 9.9, 9.10, 9.11, 9.12, 9.13).

9.3 Results

The presented ts-AUC test was applied using the features derived from statokinesigrams from Eyes-Open and Eyes-Closed acquisitions. Table 9.3 contains the obtained p-values for the two groups by the application of the ts-AUC and MMD tests. Both these tests agreed that the features derived by statokinesigrams of Eyes-Open significantly separated PS_F from PS_{NF}, contrary to those from Eyes-Closed that did not show a significant result (Table 9.3). Therefore, we will henceforth continue by presenting detailed analysis only for Eyes-Open features.

Data type	MMD result	ts-AUC result
Eyes-Open	H_0 rejected *	p-value < 0.01 *
Eyes-Closed	H_0 not rejected	p-value > 0.05

Table 9.3. The p-values obtained by the application of the ts-AUC and MMD tests on the features extracted from Eyes-Open and Eyes-Closed statokinesigrams. Features derived by Eyes-Closed statokinesigrams did not show a statistically significant result neither using ts-AUC nor MMD test. Therefore the study did not proceed to further analysis of these statokinesigrams. The statistically significant results are indicated by ' * '.

The most influential features were found to be the `VelocityY`, `VarianceY`, `AccelerationY`, `EllArea` (Confidence Ellipse area) and `MaxX` (see in Figure 9.3 their relative importance and in Figure 9.4 their mean $\pm$ standard deviation per group). Table 9.4 indicates those features that showed p-value < 0.05 and the decisions regarding statistical significance obtained after applying each of the three employed corrections. In every row of Table 9.4, values at column 1 compared one by one to values at columns 2, 3 and 4 were found **always higher**. Interestingly, although the `AccelerationY` did not show statistical significance after the MWW application (p-value > 0.05), it was found as one of the influential features by the ts-AUC test. According to Table 9.4, using the results from the three corrections with level $\alpha = 0.05$, none of the features would reject the H_0 of two-sample MWW test.

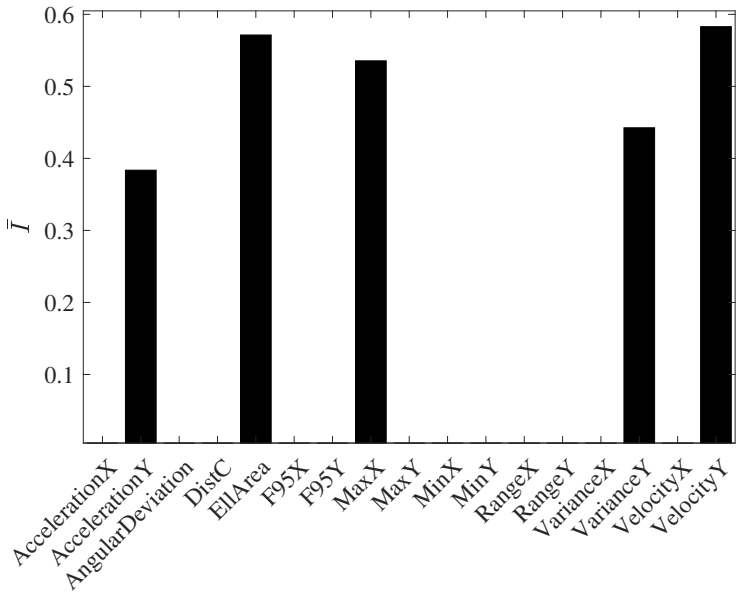


Figure 9.3. Important Features. The importance of features as estimated by applying the approach of Genuer et al. (2010) using the hyperparameters that produced the RF*.

9.3.1 Population size

As expected, the decrease of population size had an important effect to the performance of all tests. Both ts-AUC and MMD test showed similar behavior with the progressive decrease of population size. Specifically, the number of times that the fallers and non-fallers were found statistically different was gradually decreased. After 55% of population size decrease, the two groups were found significantly different in less than 50% of the cases (Figure 9.5). Univariate testing through MWW

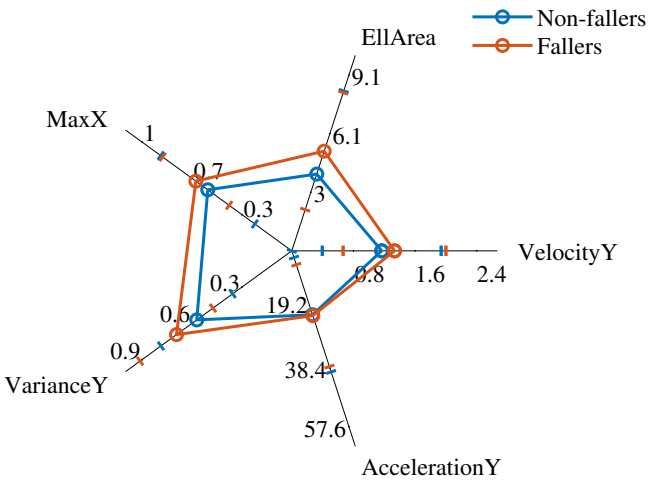


Figure 9.4. Radar chart comparing fallers and non-fallers based on the mean (o) ± standard deviation (-) of the most important features of our analysis. All six features are positively correlated with low postural control, which justifies the meaningfulness of inspecting the area of the curves in this chart. The profile of the two groups is significantly different.

followed a similar decrease. Multiple testing showed that the groups cannot be considered as statistically different(almost always).

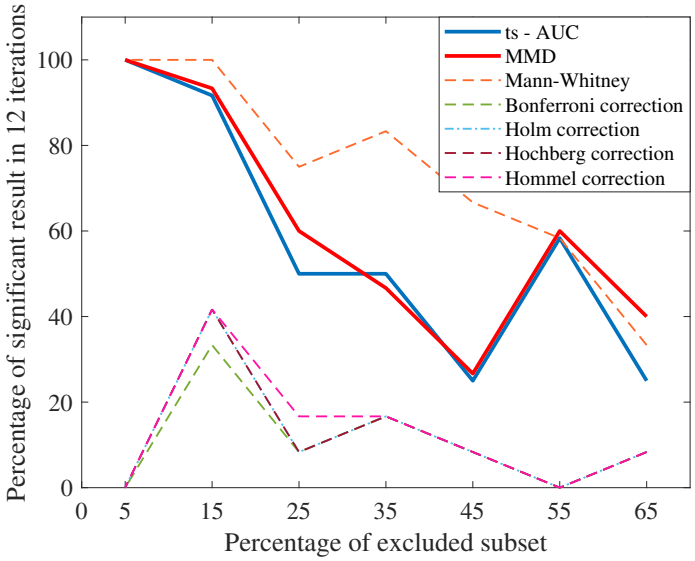


Figure 9.5. The average performance of two-sample testing approaches with smaller population. The dataset size was progressively decreased by a step of 10%. The included subset of each step was selected uniformly at random 12 times and the tests run in every iteration. We observe that ts-AUC and MMD have almost the same performance. Decreasing the population leads to lower chance of distinguishing the two groups. On the other hand, all the two-sample corrections present significantly lower performance.

	p-value before correction	*p*-value after correction			
Feature	*p*-value	Hommel	Hochberg	Holm	Bonferroni
EllArea	0.0045	0.058	0.071	0.071	0.072
VarianceY	0.006	0.092	0.11	0.11	0.12
MaxY	0.006	0.092	0.11	0.11	0.12
DistC	0.007	0.10	0.11	0.11	0.13
RangeY	0.008	0.12	0.13	0.13	0.17
VelocityY	0.009	0.24	0.33	0.36	0.50
MaxX	0.03	0.32	0.33	0.36	0.51
RangeX	0.04	0.34	0.41	0.47	0.79
VarianceX	0.04	0.36	0.41	0.47	0.82
MinY	0.04	0.41	0.41	0.47	0.87
MinX	>0.05	-	-	-	-
VelocityX	$\gg$	-	-	-	-
AccX	$\gg$	-	-	-	-
F95X	$\gg$	-	-	-	-
AccY	$\gg$	-	-	-	-
F95Y	$\gg$	-	-	-	-
AngularDev	$\gg$	-	-	-	-

Table 9.4. Significant and non-significant results of a univariate two-sample Mann-Whitney Wilcoxon (MWW) test, and the *p*-values after Hommel, Hochberg, Holm and Bonferroni corrections. After all corrections, none of the *p*-values were found lower than α level of 0.05. Therefore, none of the features can safely reject the null hypothesis at the default 5% significance level.

Regarding Figure 9.6, that shows the important role of the size proportion among the groups, the performance of ts-AUC, MMD, and multiple testing were comparable to those from Figure 9.5 (uniform decrease of the population size). However, ts-AUC and MMD exhibit a less abrupt decrease of performance. On the other hand, the gradual balancing of the sizes of the two groups, through the exclusion of non-fallers, seems to have a minor effect on the univariate MWW testing.

9.4 Discussion

The objective of this study was to introduce an easy, interpretable, and intuitive multivariate two-sample testing strategy. The particular interest of this study was to highlight the beneficial effect that this approach can have in clinical research, and particularly in the research of postural control in PS patiens. Using the proposed statistical testing approach, it was shown that: a) Different profiles between fallers and non-fallers were observed only for Eyes-Open protocol; b) The fall-prone PS patients have significantly different statokinesigram profile during quiet standing from those who are non-fallers, contrary to the classic multiple testing approach which did not agree with such a result; c) The novel multivariate two-sample testing approach (ts-AUC) showed equal performance with the state-of-the-art Maximum Mean Discrepancy (MMD) test, with the additional element of providing feature importance assessment without further analysis. d) The `VelocityY`, `VarianceY`, `AccelerationY`, `EllArea` (Confidence Ellipse area), and `MaxX`, appeared to be the most important features for distinguishing fallers and non-fallers.

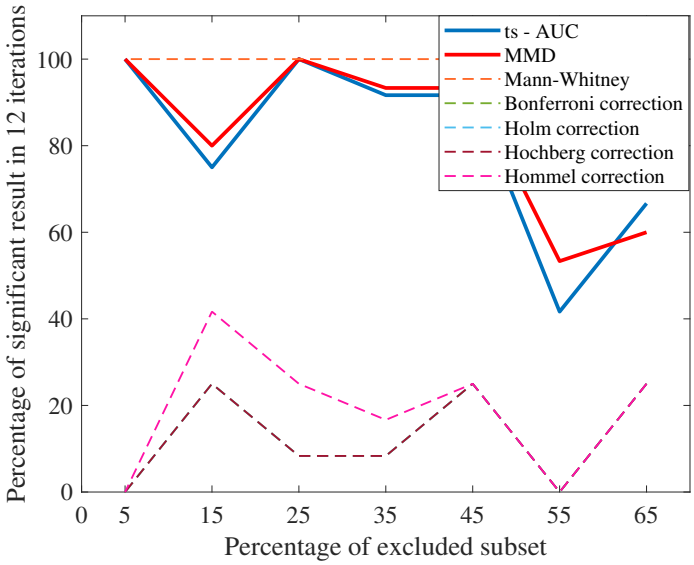


Figure 9.6. The average performance of two-sample testing approaches with smaller non-faller population. The non-fallers were progressively excluded, by a step of 10%, in order to balance the size of the two groups without excluding fallers. The included subset of each step was selected uniformly at random 12 times, all fallers were included, and the tests run in every iteration. We observe that ts-AUC and MMD have almost equal performance. Decreasing the non-faller population leads to lower chance of distinguishing the two groups. On the other hand, all the two-sample corrections present significantly lower performance.

One of the main results of this chapter is that the proposed multivariate two-sample test, the ts-AUC, and the standard statistics (usually used in clinical studies), when both applied to the dataset of PS patients lead to contradictory conclusions. The multivariate approach found fallers' and non-fallers' statokinesigram characteristics significantly different, while traditional statistics did not confirm this result. In line with previous works Feise (2002); Perneger (1998a), the applied p-value correction strategies are found to be more strict in controlling the Type I error, compared to the proposed multivariate alternative.

Researchers can always perform multiple univariate tests and not apply correction strategies (see univariate MWW results in Table 9.4, Figure 9.5, and Figure 9.6), and take the risk of having a false-positive finding. However, when modest evidence is found in relatively small populations after multiple testing, then the aforementioned false-positive probability is significantly high. The level of that risk may be controlled when some criteria are met (see Feise (2002)) considering the quality of the study, the quality of the dataset and the clinical strength of pre-set hypotheses. In exploratory studies though, some of the p-values around 0.05, whichever side they may lie on, would definitely be considered as "interesting hints", whereas concluding without thoughtful consideration from such findings should be generally avoided Wood et al. (2014). The multivariate and cross-validated approaches can decrease the aforementioned uncertainty. The proposed ts-AUC test has interesting and convenient properties: it is a test which is easy to implement and interpret, while it can be also applied to other similar multidimensional datasets.

The features included in our analysis have been used by clinical researchers in the past. Most of them were proposed as indicators of balance impairment at least once in the clinical literature (indicative references Błaszczyk et al. (2007); Mancini et al. (2012b); Melzer et al. (2004); Muir et al. (2013)). We deliberately avoided any feature engineering or transformation process, not only because that goes beyond the scope of this study, but also because we intended to focus particularly

on the merits of the newly proposed approach.

Interestingly, only the Eyes-Open acquisition allowed to significantly distinguish fallers from non-fallers in a population of PS patients. This result seems slight contradictory since PS patients exhibit increased dependency on visual sensing Rinalduzzi et al. (2015). By exploiting the advantage of the ts-AUC test that provides automatically the importance assessment of features, we found that medio-lateral movement played also a role in faller/non-faller separation of PS patients (see Figure 9.3 and Figure 9.4). The medio-lateral movement has been reported as the most discriminative element between PS patients and age-matched controls Mancini et al. (2012a) and seems that play a role in distinguishing fallers and non-fallers PS patients. However, the key-difference between fallers and non-fallers was spotted in antero-posterior movement. `VelocityY`, `VarianceY`, and `AccelerationY`, which may carry overlapping information, were found among the most influential features in the fallers/non-fallers separation. The aforementioned result is in line with previous works that reported increased antero-posterior movement of PS patients in quiet-standing conditions with eyes open Kerr et al. (2010); Korpelainen et al. (2007); Latt et al. (2009). Although many PS patients with low postural control did not manifest large posturographic areas, the confidence ellipse area (`EllArea`) was found significantly larger in fallers compared to non-fallers (Figure 9.4). However, the `EllArea` value of non-fallers was highly dispersed. Therefore larger fallers cohorts are needed in order to draw safer conclusions. The confidence ellipse area is recommended to be always considered together with antero-posterior features such as variance and velocity, in order to perform more accurate postural control assessment.

The choice of using the OOB observations as cross-validation method has two basic advantages: 1) provides faster results in the AUC maximization process, and 2) allows the final MWW test to be applied once to the whole dataset, which is more intuitive for clinicians. In cases where the population size is sufficiently large and the hypothesis of similar distributions between train and test-sets is not violated, it is expected that more classic methods such train-test split (as originally proposed in Clémençon et al. (2009)) would have given the same result (or even better; OOB prediction error results have been reported as slightly overestimated Janitza and Hornung (2018)). However, clinical datasets are usually limited in size and the aforementioned assumption about the same distribution is not always fully guaranteed. In these cases, multiple train-test splits seem more appropriate whereas they would significantly increase the testing process. OOB observations can be seen as an internal multiple train-test split (one per tree) of the RF (each observation's prediction is predicted by less than T trees) but,conveniently, the final two-sample MWW test is applied once to the whole dataset after the validation process.

Another important modification is the addition of unbiased feature importance through random permutation of OOB observations. We believe that this property is a cornerstone of the proposed approach and inline with the current clinicians' needs. While they need to know if two groups are (or are not) significantly separated, they are also interested to know the most influential features that lead to the reported result. Although the algorithm offers this convenience, we need to note that feature importance should be treated with extra care.The proposed approach tries to minimize the false conclusions concerning the importance of features when redundant features are present. According to Genuer et al. (2010), some of the collinear features (relevant to the phenomenon) will be in the final selection, and others will not. This issue is still under research and the current ts-AUC framework can integrate better solutions in the future. A general advice to clinicians can be to check for features exhibiting mutual information before the beginning of the testing process.

The features computed by the basic Romberg test have been reported as relatively inconclusive in distinguishing fallers and non-fallers, mainly due to the lack of realistic conditions of fall Palmieri et al. (2002). The available patients' dataset, with its relatively 'marginal' separation between fallers and non-fallers (see Table 9.4), can be considered as an ideal dataset in order to check the perform-

ance of the newly proposed approach. We consider MMD algorithm as the gold-standard method in terms of separability of the two groups. The fact that ts-AUC shows similar performance to that of MMD is very important, especially if we think that the proposed ts-AUC can also provide additional information about the most influential features without the need of any supplementary (meta-)analysis. Therefore, it would be fare to say that ts-AUC is competitive in terms of performance, while also boosting the interpretability of the result for the convenience of clinicians.

Interestingly, the decrease of the overall population and the gradual balancing between the groups of fallers and non-faller, showed that the proposed test is less conservative than the multiple testing process (with corrections). Exploratory studies, where a hypothesis about the structure of the dataset is not strictly defined in advance, could benefit from such multivariate approaches.

Comparing the results of the two population reduction schemes, i.e. the uniform reduction of the population versus the reduction of non-fallers (the larger group), we observe that all the statistical tests performed slightly worse in the former case. This was an expected result since fallers were only 24 out of the 123 available PS patients, and thus decreasing the size of that group made the fallers heavily underrepresented in the produced subsample.

Limitations. The first limitation of this study is the lack of sufficient evidence about the reasons behind falls. The basic Romberg test has been reported to be an insufficient protocol to provide such physiological information Palmieri et al. (2002); Swanenburg et al. (2010). Previous studies proposed richer protocols (including multi-tasking or use of foam surfaces Chagdes et al. (2009); Melzer et al. (2004); Swanenburg et al. (2010)) for postural control assessment of fragile individuals such as PS patients. Undoubtedly, such protocols can have beneficial effect to the faller/non-faller classification, as well as to the impairment assessment of patients (visual, vestibular, somatosensor, nervous system). Yet, among the objectives of this work was to show that basic Romberg test does contain fall risk-related information, whose extraction and full exploitation is largely up to the adequacy of the employed statistical analytics.

It is worth noting that there is always some uncertainty in what patients report as their recent fall experience. Participants who were asked about previous falls might confabulate without a conscious intention to deceive (recall bias). Therefore, some of the non-fallers might be mistakenly labeled as non-fallers. Machine learning algorithms are usually robust to the presence of such noise. Besides, in medical studies the sample size is most usually small, as in ours, and it is required to prepare carefully the population to study. Therefore, this kind of noise is usually minor since patients are actually interviewed by medical experts who can identify subjects that could bring uncertainty to the analysis and exclude them from the sample.

In extreme cases of imbalanced datasets with many negative values and few positive ones, other metrics rather than the AUC, such as the precision-recall (PR) curve, the F_1 score, or the area under the PR curve, could be more appropriate in order to prevent overfitting Davis and Goadrich (2006) (AUC still remains robust to imbalanced datasets). We decided to keep the AUC criterion, which is the one initially proposed by Clémençon et al. (2009), in order to fulfill one of our main objectives: to propose the algorithm as understandable, interpretable and easy-to-implement as possible. In return, as it has been already mentioned, we controlled the leaf size (LS) and the number of features (M) in the optimization procedure, and we applied cross-validation in each resulting case.

The use of Wii Balance Board (WBB) as a force platform during the acquisition protocol, is another mentionable limitation. The reliability of the WBB as a medical examination tool has been previously questioned Pagnacco et al. (2011). Basic reported drawbacks were: a) the modest agreement with laboratory grade force platforms, b) the lower signal to noise ratio in its recording, and c) the irregular sampling rate Castelli et al. (2015). We state that we are perfectly aware of the aforementioned limitations. However, Wii Balance Board presents an increasing popularity in pos-

turography studies as a valid tool for assessing standing balance Clark et al. (2010); Leach et al. (2014). It is an inexpensive piece of equipment and hence seems ideal for applications that intend to provide a quick and low-cost first scan of individuals with certain possibility of postural control loss. In addition, recent works Audiffren and Contal (2016); Leach et al. (2014) showed that a careful pre-processing can mitigate some of its aforementioned drawbacks.

9.5 Conclusion

In this chapter we showed that using the proposed ts-AUC two-sample test, which is a based on AUC maximization, faller and non-faller patients who suffer from Parkinsonian syndromes (PS) can actually be distinguished by examining posturographic features that are derived following the basic Romberg protocol. This novel approach was also able to reveal the posturographic features that are significantly different between the two groups (more discriminative). We confirmed that a fall-prone PS patient may manifest wider and more abrupt antero-posterior oscillations and larger posturographic areas compared to a non-faller. This separation appeared statistically less detectable when using more traditional approaches such as multiple testing. Interestingly, the above results were observed only in statokinesigrams derived by the Eyes-Open protocol. The results of our study highlighted that new multivariate methods based on machine learning, such as the ts-AUC test, can play an important role in evaluating the usefulness of simple and inexpensive acquisition protocols as well as the extracted posturographic features. We plan to generalize the current framework. Nevertheless, any extension should investigate the statistical metrics that would be theoretically suitable to be used as optimization criteria.

9.5.1 Additional results in simulated datasets

We conducted additional experiments to test and compare the performance of ts-AUC using simulated datasets and we provide it as a supplement to the analysis on the real use-case of we studied in the main text. The figures appearing below compare ts-AUC with MMD and a multiple testing procedure with p-value correction.

Simulated data. We created datasets by mixing two independent Gaussian groups. For each dataset we pick:

- the population size ($N = 100$ or 200);
- the proportion of the two groups forming the population (50%/50%, 70%/30%, or 90%/10%);
- the number of dimensions (10, 20, or 30) mimicking the amount of variables that a usual clinical study may have;
- 2/3rd of those dimensions had no difference between the two groups by design (generated using exactly the same average and standard deviation).
- the remaining 1/3rd had a progressively increasing difference in their average (x-axis in all figures below).

We run the test 20 times per each combination case. We compared the performance of ts-AUC, MMD, and multiple testing with p-value correction (We only mention Hommel and Hochberg for lisibility reasons as well as due to their power superiority compared to the others), keeping the percentage of significant results that each test acquired (y-axis). In all generated cases of non-extreme

proportions (50/50, 70/30) between groups' sizes (Figures 9.7, 9.8, 9.9 and 9.10), ts-AUC and MMD present similar behavior, and they were always superior to multiple testing approaches in detecting the difference between the two groups. In cases of highly imbalanced groups (see Figures 9.11 and 9.12), there is no clear superiority of any method; all methods have increased Type I errors since the generation of the minority group is not reliable.

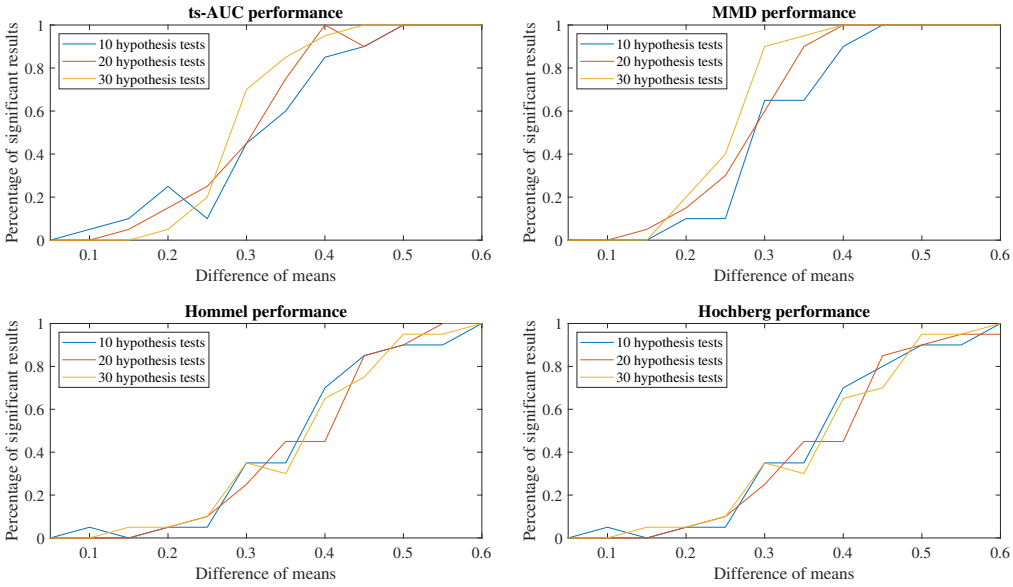


Figure 9.7. The average performance of two-sample testing approaches in simulated datasets with class balance 50/50, and 10, 20, or 30 features. We observe that ts-AUC and MMD have almost the same performance and always superior to the multiple testing strategies.

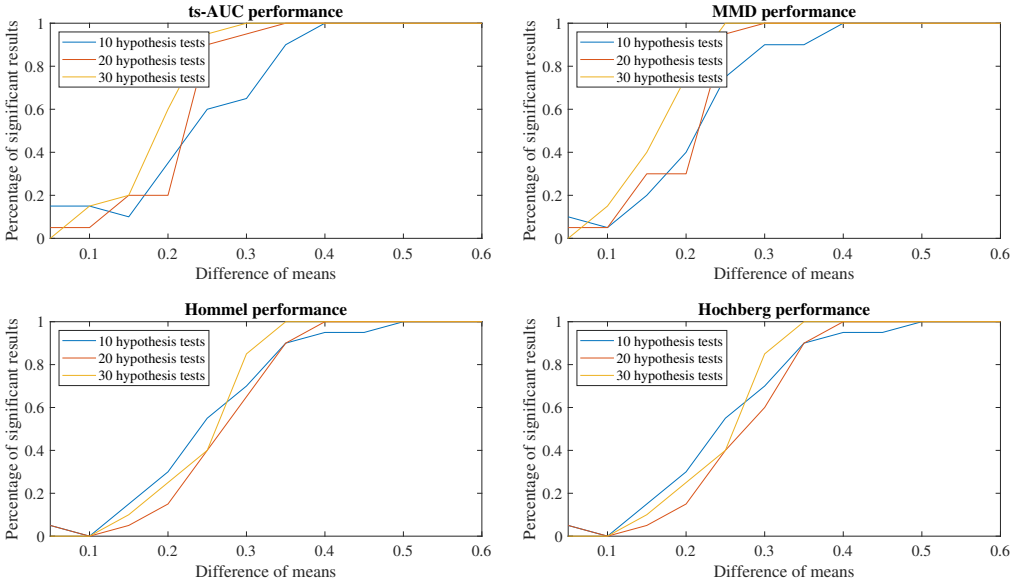


Figure 9.8. We observe also that ts-AUC and MMD have almost the same performance and always superior to the multiple testing strategies, especially for the cases of >10 hypothesis tests.

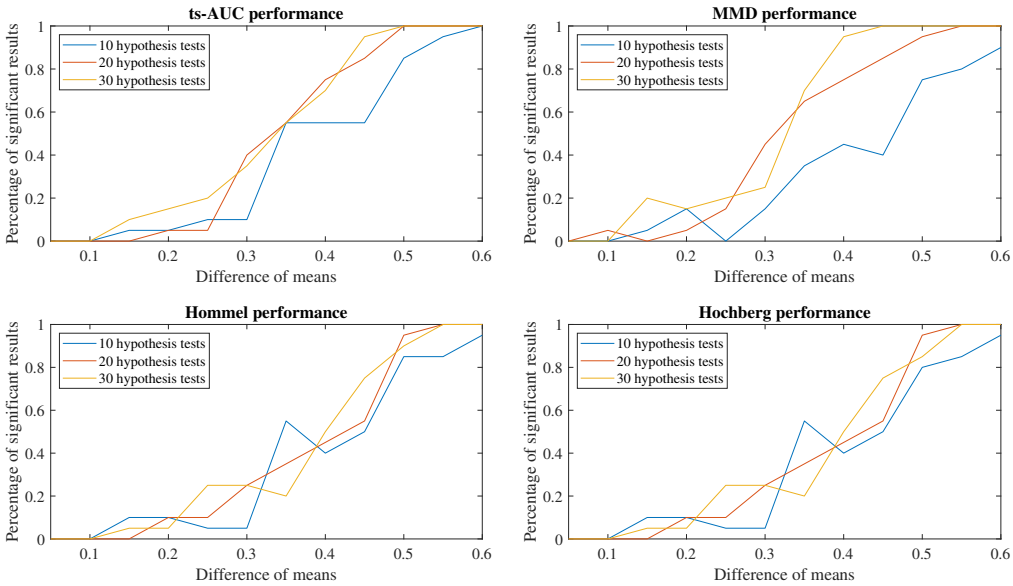


Figure 9.9. In this setting, we observe that ts-AUC and MMD have almost the same performance. Introducing class imbalance reduces the chance of distinguishing the two groups mainly due to the low representation of the minor group. The two-sample corrections are affected more and present significantly lower performance than in the balanced case.

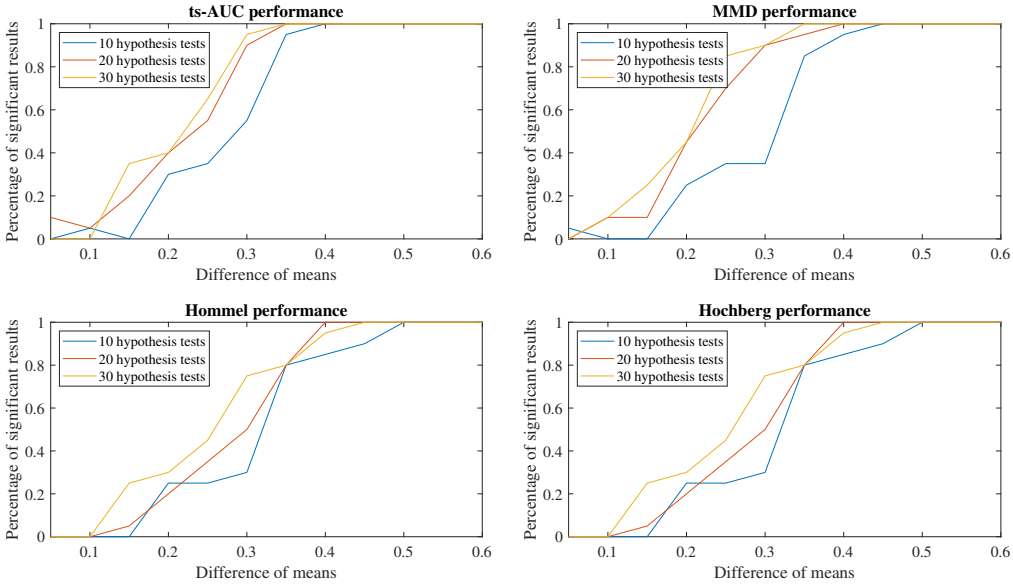


Figure 9.10. In this setting, we observe that ts-AUC and MMD have almost the same performance (ts-AUC is slightly better in case of 10 hypothesis test). Some Type I errors might be present in both multivariate tests.

9.5.2 Feature importance and population

We created two independent Gaussian groups of:

- various total populations ($N = 50, 100, 150, 200$);
- 50%/50% balance between groups;

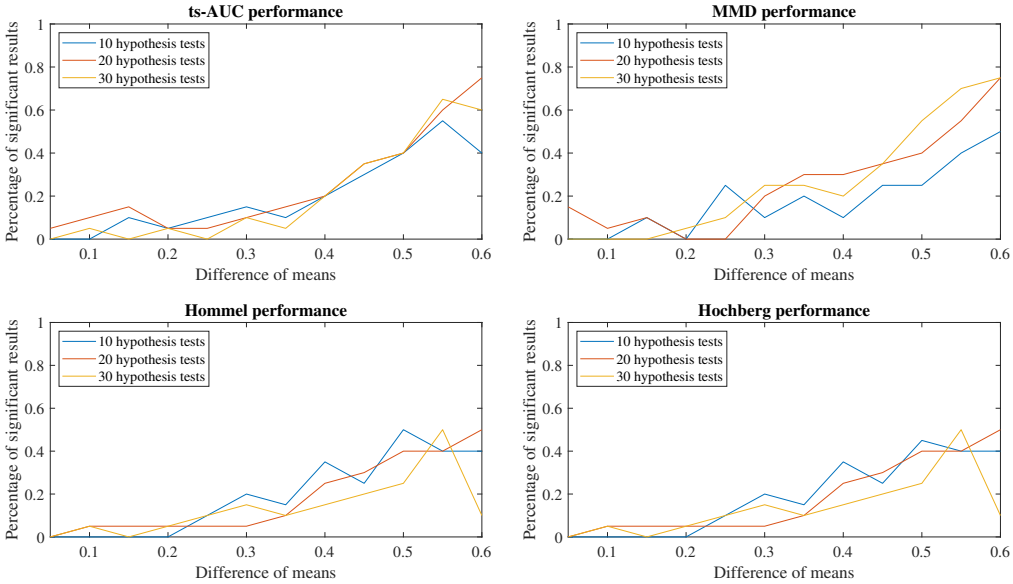


Figure 9.11. In this setting, we observe that all approaches have almost the same performance. For mean difference >0.5, it seems that the two multivariate approaches, ts-AUC and MMD, begin to have superior performances. However, they also tend to have higher Type I errors. Generally, one of the groups is extremely small (size of 10) for a reliably distinguished distributions at simulation process.

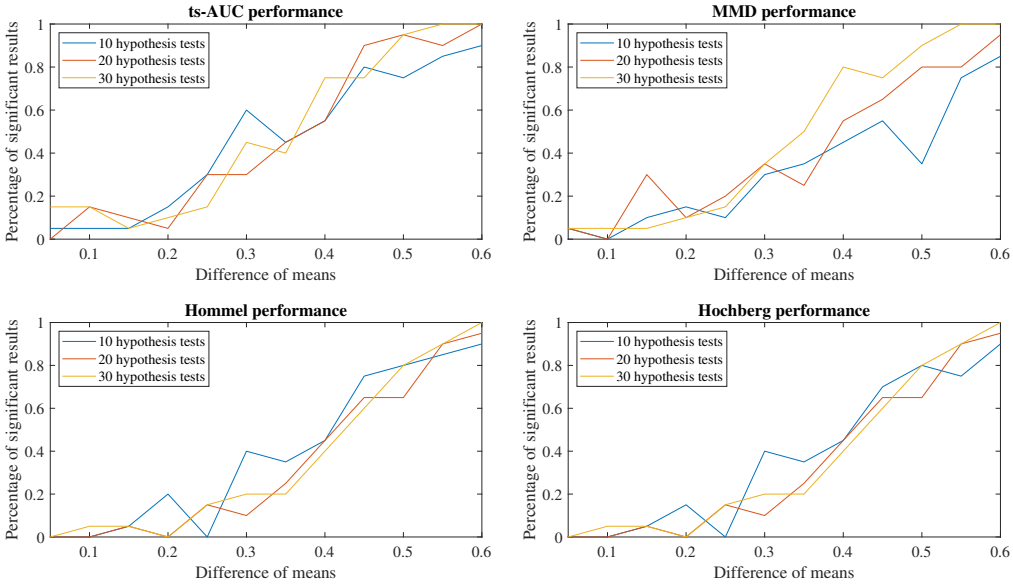


Figure 9.12. In this setting, we observe that in this special case, still ts-AUC has the best overall performance. We now see more reasonable results due to the fact that the minority group (now size of 20) can marginally have a reliably distinguished distributions at simulation process. However, TYPE I errors are still present.

- 30 dimensions (features);
- three quarters of those features (no. 1-22) had no difference between the two groups by design (all generated using exactly the same average and standard deviation - $\mathcal{N}(0, 1)$);
- the remaining one quarter (no. 23-30) were generated by $\mathcal{N}(0.9, 1)$;

- no colinearities between features.

By design, the features 23-30 are significantly different between the two generated groups. We performed 10 runs of the algorithm for every population. Indeed, the feature importance element of the algorithm performed effectively and found as more important the features that by design were more different between the two groups (see Figure 9.13). The proposed algorithms almost always selected as important elements only those which had by default significant difference between the groups. More details about the limitations of the current feature importance algorithm can be found in the Limitations part at the end of the Discussion section.

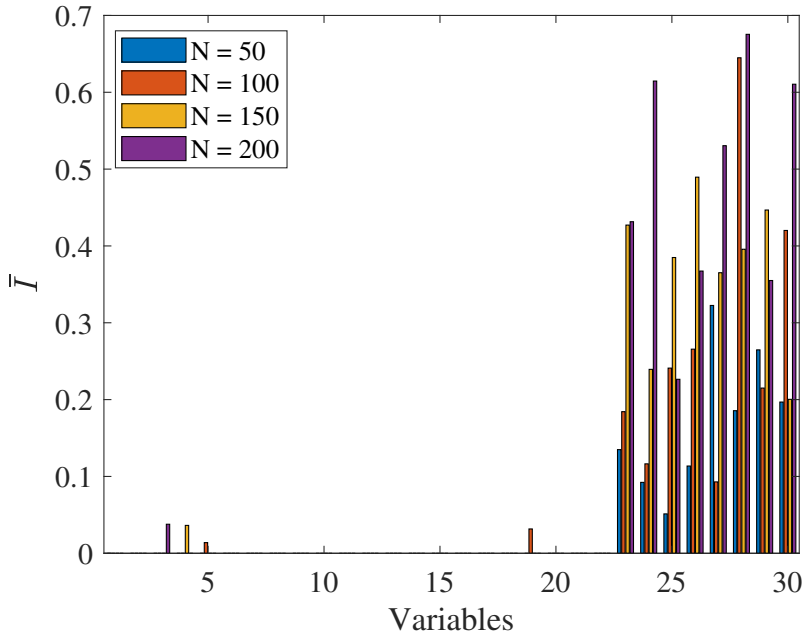


Figure 9.13. Features 23-30 are by design significantly different between the two generated groups. We observe that ts-AUC detects effectively the important elements in all populations.

10 | A Generative Model for the Postural Control

Abstract. Falls are a major concern of public health, particularly for older adults, as the consequences of falls include serious injuries and death. Therefore, the understanding and evaluation of postural control is considered key, as its deterioration is an important risk factor predisposing to falls. In this work we introduce a new Langevin-based model, local recall, that integrates the information from both the center of pressure (CoP) and the center of Mass (CoM) trajectories, and compare its accuracy to a previously proposed model that only uses the CoP. Nine healthy young participants were studied under quiet bipedal standing conditions with eyes either open or closed, while standing on either a rigid surface or a foam. We show that the local recall model produces significantly more accurate prediction than its counterpart, regardless of the eyes and surface conditions, and we replicate these results using another publicly available human dataset. Additionally, we show that parameters estimated using the local recall model are correlated with the quality of postural control, providing a promising method to evaluate static balance. These results suggest that this approach might be interesting to further extend our understanding of the underlying mechanisms of postural control in quiet stance.

Contents

10.1 Introduction

In our ageing societies, falls are a major concern of public health cdc (2017). This is especially true for older adults, as the consequences of falls are more severe, including serious injuries and death Sterling et al. (2001), and their prevalence is high, as each year more than a third of population 65 years-old and older faces a fall Tinetti (2003b). While strategies of prevention and rehabilitation have been shown to be effective to reduce those falls and their consequences Rubenstein (2006); Van Diest et al. (2013), they rely on an early detection and accurate characterization of the individual balance related deficiencies. In particular, the evaluation of postural control – the ability to maintain equilibrium and orientation in a gravitational environment Horak (1987) – is considered key to this end Perrin et al. (1997) as its deterioration is an important risk factor predisposing to falls Rubenstein (2006).

Postural control results from the complex synergy between the central nervous system, the musculoskeletal system and the sensory entries (visual, somatosensory and vestibular) Horak (2006); Kurz et al. (2013); Loram (2015); Peterka (2002). A common approach to evaluate postural control is to use protocols that inhibit visual (such as the Romberg test Khasnis et al. (2003)) or proprioceptive feedbacks (using e.g. foam Patel et al. (2008)) while recording the position of the center of pressure (CoP) – the point of application of the ground reaction forces resultant under the feet Lafond et al. (2004) – over time using a force platform. The recorded two-dimensional signal, which includes both the medio-lateral (ML) and antero-posterior (AP) axes, can be used to analyse the neuromuscular control involved, and in particular the adjustments performed by the individual to maintain balance, i.e. to keep the projection of the center of mass (CoM) inside the base of support Baratto et al. (2002); Winter et al. (1998).

In previous works, multiple descriptors derived from the CoP have been shown to capture discriminatory characteristics of postural control Baratto et al. (2002); Caron et al. (2000); Corriveau et al. (2004). More precisely, they were shown to present statistically significant different values among distinct populations such as older adults fallers, athletes, or individuals with neurological disorders such as Parkinson's disease Mitchell et al. (1995); Nicolaï and Audiffren (2018). Those descriptors can be general statistics of the signals, such as mean velocity or sway density Prieto et al. (1996), or parameters derived from dynamic models Collins and De Luca (1993); Hernandez et al.

(2015); Peterka (2002). A significant benefit of the later approach is that it enables an interpretable parametrization of trajectories that arises directly from the formulation of the dynamic model.

Interestingly, several of these aforementioned models have assumed the presence of randomness in the CoP trajectory, due to either self-induced perturbations of postural control or external perturbations such as respiration Bottaro et al. (2005), as well as the inaccuracy of the sensorimotor system. Consequently, these previous studies have proposed to model the CoP signal as a stochastic process Collins and De Luca (1993); Newell et al. (1997). For instance, it has been suggested that the CoP displays a mean quadratic displacement similar to the one of a fractional brownian motion with two regimes Collins and De Luca (1993). Other works have proposed to model the CoP dynamic using Langevin differential equations Bosek et al. (2004, 2005); Lauk et al. (1999); Tawaki and Murakami (2019). This model has shown to be promising to reproduce intrinsic characteristics of the trajectory Tawaki and Murakami (2019). In this setting, the acceleration of the CoP is expressed as the combination of several of the following forces: a spring restoring force, also called recall, a damping force and a Brownian motion.

Possible interpretations have been proposed for these forces. For instance, the recall force has been used to express the corrective force acting on the CoP to pull it back towards a reference position Hernandez et al. (2015); Lauk et al. (1999).

This is in line with previous works that have advocated for the existence of a mechanism that produces a corrective ankle joint moment, which can be modeled as a spring restoring force, eventually damped Peterka (2002); Winter et al. (1998, 2001). Since the parameters of each force can be estimated using e.g. ordinary least-square method applied on transformations of the signal, such as the mean squared displacement Bosek et al. (2004), or directly on the CoP signal Tawaki and Murakami (2019), it is possible to evaluate the relative importance of each force, hence giving insights about the characteristics of balance control. It has been claimed for instance that this model enables the evaluation of individual stiffness Lauk et al. (1999).

However, we argue that these Langevin-based models can be significantly improved by including the Center of Mass (CoM) as part of the system. Indeed, one popular hypothesis states that the CoM trajectory operates as a moving reference position, from which any deviation results in the activation of appropriate restoring forces Gatev et al. (1999); Zatsiorsky and Duarte (1999). This assumption has been strengthened by previous studies which have shown that during quiet stance, the CoP oscillates in phase with the CoM with higher amplitudes Winter et al. (1998). Moreover, this hypothesis has been successfully applied to continuous linear feedback controllers with time delay system such as PID (Proportional, Integral, Derivative) systems to model the control of body deviations Mahboobin et al. (2007); Masani et al. (2006); Peterka (2002).

In line with these results, in the present study we introduce a new Langevin model that includes a recall force pulling the CoP toward the CoM, in addition to a damping force and a Brownian noise. Our experiments show that this model significantly improves the quality of the CoP trajectory predictions, compared to a commonly used Langevin model, including when subjects' vision and/or standing surface were manipulated. Additionally, we show that the same results can be obtained on another publicly available dataset of postural control dos Santos et al. (2017). We also used this new model to estimate the relative importance of each force, and our results show that these parameters can be used to differentiate between distinct populations and experimental conditions, highlighting their possible use to improve the understanding of several aspects of postural control. Overall, the present study supports the hypothesis that the CoP dynamic is intrinsically and deeply intertwined with the CoM dynamic, and that the Langevin model has the potential to quantify interesting components of postural control and can greatly benefit from encoding joint dynamic of the CoP and CoM instead of their marginal behavior.

10.2 Method

10.2.1 Participants and protocol

In this study we analyzed our model using two different populations. For the first population (hereafter referred to as population 1), 9 healthy young participants were recruited specifically for this study (age: 27.6 $\pm$ 7.1 years, weight: 73.0 $\pm$ 6.5 kg, height: 170.0 $\pm$ 10.1 cm, three females). All participants were right-hand dominant with normal or corrected to normal vision. All participants signed an informed consent document approved the 22 July 2020 by the IRB of the Fribourg University, Switzerland, ref. 583R3.

Participants were asked to stand still with feet at pelvis width, arms laying at the side. For each acquisition, this quiet stance was recorded for 50 seconds, using a force platform and a kinematic system. Each participant was recorded twice for each possible combination of the following conditions: eyes open or closed, and standing on a surface that was either rigid or foam. During trials with eyes open, participants were asked to fix a target which was located at eyes height, at two meters distance. Trials were acquired in blocks of two consecutive recordings, in order to reduce confounding factors, such as fatigue or learning Hernandez et al. (2015). In between blocks, subjects were allowed to rest by sitting or walking around.

We replicated our results by using the public dataset dos Santos et al. (2017) (hereafter referred to as population 2), which contains three-dimensional kinematics and the ground reaction forces of 49 subjects (27 young individuals – 15 males, 12 females – between 18 and 40 years old; 22 older adults – 11 males and 11 females – 60 years old or older). The database contains 588 recordings in total, among which 17 were removed as their kinematics time series were missing. All subjects were recorded in similar conditions as the first population, and both subjects' vision and the standing surface were identically manipulated.

10.2.2 Hardware

For our study, CoP data were collected using a ground-level six-channel force plateform (AccuSway, AMTI, Watertown, MA, USA), which sampled the three-dimensional ground reaction forces and moments at 100 Hz. A poster was used to provide a 5-cm fixation target that was displayed approximately two meters in front of the participant, at eye level, during eyes open conditions. In order to standardize the shoe–platform interface, participants were recorded while wearing standardized socks. Kinematic data were collected using an OptiTrack system (NaturalPoint, Corvallis, OR, USA) at a sampling rate of 100 Hz using 18 cameras. Each participant wore a full body suit, on which markers were placed to track the position of key anatomical locations, which were used to compute the position and trajectory of the CoM during the recording. More specifically, markers were positioned following the model defined in Lafond (2003); Zatsiorsky and Zaciorskij (2002). The detailed position of the markers can be found in Table 10.1.

10.2.3 Data preprocessing

Data from the force platform and the Optitrack system were collected and synchronized using Motive (NaturalPoint, Corvallis, OR, USA). Data preprocessing and analysis software were written using Python (v3.7, Python Software Foundation, OR, USA). Raw force platform data were processed with a fourth-order, zero-lag, low-pass Butterworth filter with a 10 Hz cutoff frequency, in accordance to Hernandez et al. (2015). CoP position was calculated with the usual formula Sandholm et al. (2009):

$$\mathrm{CoP}_x = \frac{-F_x c - M_y}{F_z} \quad \text{and} \quad \mathrm{CoP}_y = \frac{-F_y c + M_x}{F_z} \,,$$

Body part	Markers positions
Head	Vertex, Midpoints between gonions
Trunk	Acromions, Xyphoid, Antero-superior illiac splines
Arms	Acromions, Elbow joint centers, Styloid processes, and Distals of the tips of the third metacarpes
Legs	Greater trochanters, Knee joint centers, Posteriors of calcanei and Tips of the second toes

Table 10.1. Marker positions used for CoM tracking, sorted by body part.

where CoP_x (respectively CoP_y), F_x (resp. F_y), M_x (resp. M_y) denote the coordinates of the CoP, the ground reaction forces and the moments on the medio-lateral, resp. antero-posterior axis, F_z denotes the ground reaction force coordinates in the vertical axis, and c is the calibration parameter of the force platform. The resulting CoP trajectory was then centered.

Similarly, the COM trajectory was derived from the markers positions using the mass ratio coefficient defined in (Lafond, 2003, Table 3.II). The resulting three dimensional trajectory was then projected to the ground plane, centered, and processed with a fourth-order, zero-lag, low-pass Butterworth filter with a 10 Hz cutoff frequency. Finally, both the CoP and CoM trajectory were resampled at 20 Hz, corresponding to a Nyquist frequency of 10 Hz. Example of the resulting trajectories can be found in Figure 10.1.

10.2.4 Mathematical model: Local Recall

As noted by previous works, the behavior of the CoP trajectory shares important characteristics with a Wiener process Collins and De Luca (1993). However, representing the CoP by a Brownian motion implies that the CoP would exit the base of support in finite time, leading to a fall, which is in contradiction with the purpose of postural control. Therefore, previous studies have built on this remark by formulating the system as a Langevin model with additional forces, such as a damping or a spring restorative force Hernandez et al. (2015). In these works, the reference position for the spring restorative force is assumed to be static, and equal to the center of the base of support Lauk et al. (1999), or piecewise constant to model the shifting of weight between the feet Hernandez et al. (2015).

However, it has been observed that the CoP tends to oscillate around the CoM, instead of a fixed central point. Consequently, in this study, we are interested in studying and evaluating the following model, called local recall, where the CoP is assumed to be solution of the stochastic differential equation:

$$d\mathscr{V}_t^{\mathrm{CoP}} = \left[\Lambda \underbrace{(\mathrm{CoM}_t - \mathrm{CoP}_t)}_{\text{Recall}} + \Gamma \underbrace{(-\mathscr{V}_t^{\mathrm{CoP}})}_{\text{Damping}} \right] dt + \Sigma \underbrace{dB_t}_{\text{Perturbations}} \tag{10.2.1}$$

where CoM_t, CoP_t and $\mathscr{V}_t^{\mathrm{CoP}}$ are respectively the two dimensional coordinates of the CoM, the CoP and the velocity of the CoP at time t, Λ, Γ, $\Sigma \in \mathbb{R}^{2\times 2}$ are the coefficients matrices that characterize respectively the recall, damping and perturbations of the dynamic and B_t is a two dimensional Wiener process. Note that (10.2.1) is similar to the classical Langevin equation.

Also, (10.2.1) simultaneously defines the dynamic of the CoP along the ML and the AP axes. The resulting dynamics in each axis can significantly differ, a well known phenomenon in postural control Baratto et al. (2002). In this model, the ML and AP dynamics are assumed independent, thus

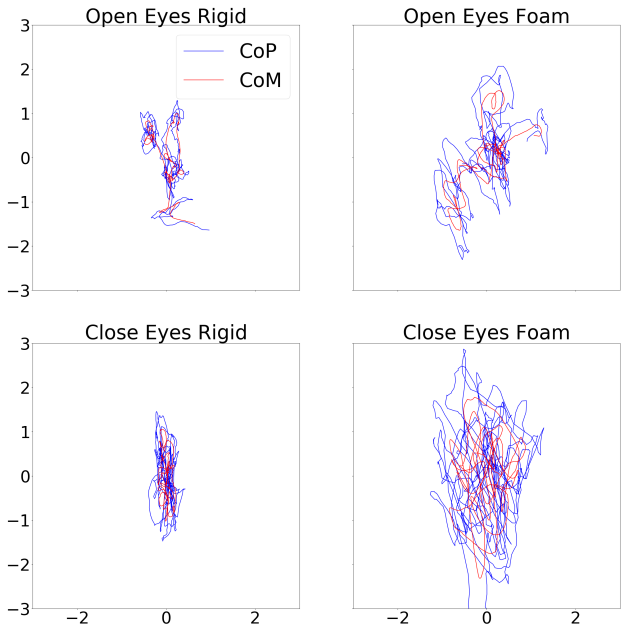


Figure 10.1. Representative CoP / CoM trajectories. The figure shows the CoP excursion (blue) and the CoM excursion (red) from a representative participant with either eyes open (top) or eyes closed (bottom) on a rigid surface (left) and on a foam (right).

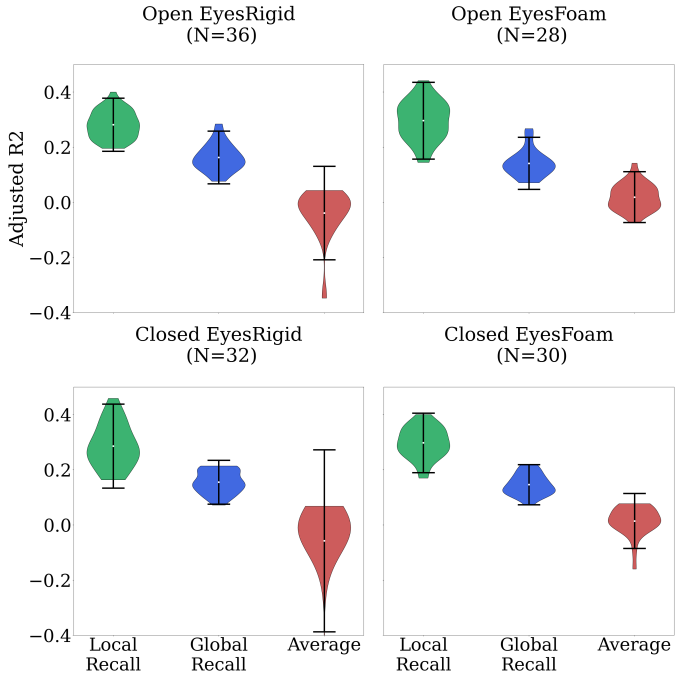
Λ, Γ and Σ are diagonal. Therefore we can write $\Lambda = \begin{pmatrix} \Lambda_{\text{ML}} & 0 \\ 0 & \Lambda_{\text{AP}} \end{pmatrix}$ where Λ_{ML} and Λ_{AP} represent the components of the local recall force applying respectively on the ML and AP axis. Γ_{ML}, Γ_{AP}, Σ_{ML} and Σ_{AP} are defined similarly.

10.2.5 Parameters estimation

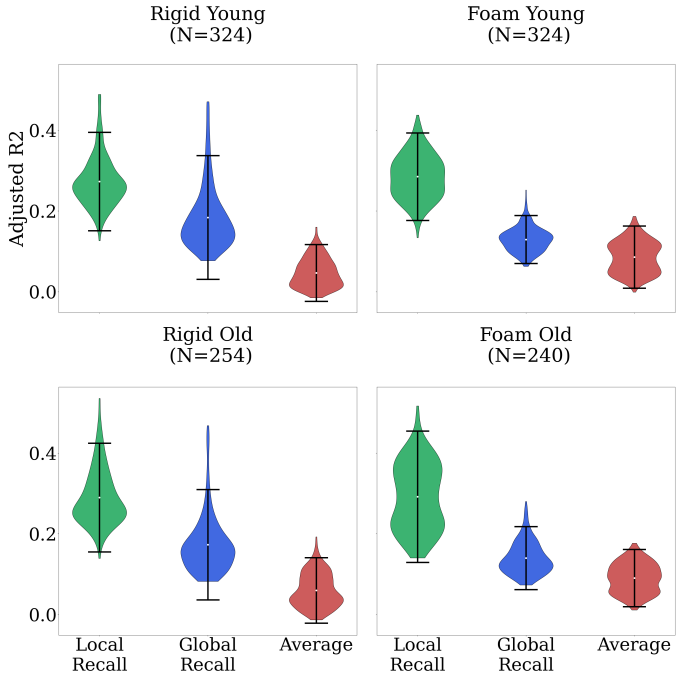
Estimating the parameters Λ, Γ, Σ in (10.2.1) is key to the analysis of this model. Indeed, different values of each parameters will result in significantly different trajectories, and we do not assume that every individual will share identical dynamics. This is particularly important as the second population studied in this work includes both older adults that have fallen multiple times and healthy young individuals dos Santos et al. (2017), two groups that have been shown to have different postural controls Alexander et al. (1992); Nicolaï and Audiffren (2018). Therefore, in order the assess the relevance of the local recall model, the task is double: first, the parameters of the model are estimated for each trajectory, using the recordings; and then the predictions of the model using these parameters are compared to the observed dynamic.

A significant difficulty regarding the parameter estimation is that while the model defined by (10.2.1) is continuous, the CoP and CoM trajectory are only observed at constant discrete time interval Δ_s (here $\Delta_s = 0.05$ s after resampling). To address this issue, we approximate the dynamic of the discrete trajectory of the CoP using (10.2.1) as follows:

$$\begin{cases} \text{CoP}_{t+\Delta_s} \approx \text{CoP}_t + \Delta_s \mathscr{V}_t^{\text{CoP}} \\ \mathscr{V}_{t+\Delta_s}^{\text{CoP}} \approx \mathscr{V}_t^{\text{CoP}} + \mathscr{N}\left(\Delta_s \mu_{\mathbf{t}}, \Delta_s \Sigma^2\right) \end{cases} \tag{10.2.2}$$



(a) Values of the adjusted R2 for population 1, for each of the four recording conditions: open eyes (top), closed eyes (bottom), rigid surface (left) and foam (right). ML and AP trajectories are jointly considered.



(b) Values of the adjusted R2 for population 2. Four subgroups are considered : Young age group (top), Older adults group (bottom), rigid surface (left) and foam (right). Open eyes and closed eyes conditions, as well as ML and AP trajectories, are jointly considered.

Figure 10.2. Distribution of values of the adjusted R2 for the local recall model (green), the global recall model (blue) and the average recall model (red) on both populations. The whiskers indicate the 95% confidence interval. In every case, the adjusted R2 values for the local recall model are significantly larger than for the global recall model.

with

$$\mu_{\mathbf{t}} = \Lambda\left(\mathrm{CoM}_t - \mathrm{CoP}_t\right) + \Gamma(-\mathscr{V}_t^{\mathrm{CoP}}) \tag{10.2.3}$$

Since CoM_t, CoP_t, and $\mathscr{V}_t^{\mathrm{CoP}}$ are known from the recording, (10.2.2) and (10.2.3) define a linear model, where the unknown parameters are Λ and Γ. We use the Ordinary Least Square (OLS) method to estimate their respective values. More precisely, using the independence of the ML and AP dynamics,

$$\begin{pmatrix}\Lambda_{\mathrm{ML}} \\ \Gamma_{\mathrm{ML}}\end{pmatrix} = \mathbf{F}_{\mathrm{ML}}^{\dagger}\mathbf{A}_{\mathrm{ML}}, \tag{10.2.4}$$

where † denotes the Moore-Penrose pseudo-inverse,

$$\mathbf{F}_{\mathrm{ML}} = \Delta_s \begin{pmatrix} \mathrm{CoM}_{\Delta_s,\mathrm{ML}} - \mathrm{CoP}_{\Delta_s,\mathrm{ML}} & -\mathscr{V}^{\mathrm{CoP}}_{\Delta_s,\mathrm{ML}} \\ \cdots & \cdots \\ \mathrm{CoM}_{n\Delta_s,\mathrm{ML}} - \mathrm{CoP}_{n\Delta_s,\mathrm{ML}} & -\mathscr{V}^{\mathrm{CoP}}_{n\Delta_s,\mathrm{ML}} \end{pmatrix},$$

is the force matrix applied to the CoP, and

$$\mathbf{A}_{\mathrm{ML}} = \begin{pmatrix} \mathscr{V}^{\mathrm{CoP}}_{2\Delta_s,\mathrm{ML}} - \mathscr{V}^{\mathrm{CoP}}_{\Delta_s,\mathrm{ML}} \\ \cdots \\ \mathscr{V}^{\mathrm{CoP}}_{n\Delta_s,\mathrm{ML}} - \mathscr{V}^{\mathrm{CoP}}_{(n-1)\Delta_s,\mathrm{ML}}. \end{pmatrix},$$

is the vector of observed speed variations. Once Λ and Γ have been computed, we are also interested in estimating Σ, since this coefficient drives the perturbation force in the Langevin model (10.2.1). We estimate Σ as the unique positive square root of the empirical variance of the residuals divided by the sampling interval, that is :

$$\Sigma_{\mathrm{ML}} = \sqrt{\frac{1}{n\Delta_s}\sum_{t=\Delta_s}^{n\Delta_s}(\mathbf{R}_{t,\mathrm{ML}} - \overline{\mathbf{R}}_{\mathrm{ML}})^2}$$

where

$$\widehat{\mathbf{A}}_{\mathrm{ML}} = \mathbf{F}_{\mathrm{ML}}\begin{pmatrix}\Lambda_{\mathrm{ML}} \\ \Gamma_{\mathrm{ML}}\end{pmatrix}$$

is the predicted speed variation matrix,

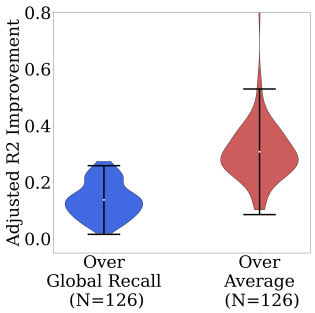
$$\mathbf{R}_t = \mathbf{A}_{t,\mathrm{ML}} - \widehat{\mathbf{A}}_{t,\mathrm{ML}}$$

is the residual of the model at time t, and $\overline{\mathbf{R}}_{\mathrm{ML}}$ is the average value of the residuals. The same process can be repeated to obtain Λ_{AP}, Γ_{AP} and Σ_{AP}.

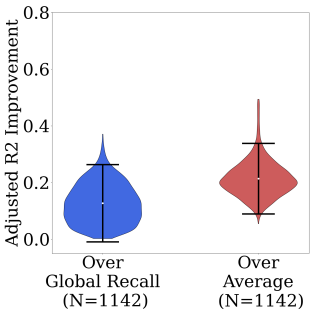
10.2.6 Model analysis

All statistical analysis were performed following the recommendations of Cumming (2014). When reported, p-values where obtained using Mann-Whitney U-test with Bonferroni correction Feise (2002), and 95% confidence intervals for estimators were obtained using the 1.96 standard deviation half width.

Performance evaluation. To evaluate the performance of the model, we proceeded as follows. For any given trajectory τ of length n, recall that $\mathbf{A}_t$ and $\widehat{\mathbf{A}}_t$ denote respectively the observed speed change of the CoP at time t and the expected speed change predicted by the linear model at time

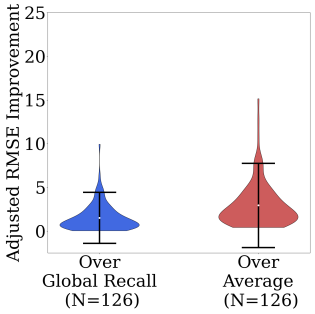


(a) Population 1.

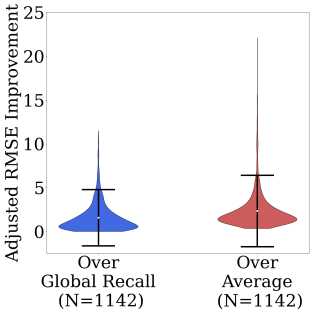


(b) Population 2.

Figure 10.3. Difference of the adjusted R2 between the local recall model minus respectively the global recall model (blue) and the average model (red) for each trajectory of both populations (left: population 1, right: population 2). The whiskers indicate the 95% confidence intervals. For every trajectory, the adjusted R2 difference is positive, highlighting the fact that the local recall model consistently produces better predictions.



(a) Population 1.



(b) Population 2.

Figure 10.4. Difference of the RMSE between respectively the global recall model (blue) and the average model (red) minus the local recall model, for each trajectory of both populations (left: population 1, right: population 2). The whiskers indicate the 95% confidence intervals. For every trajectory, the RMSE difference is positive, indicating a lower RMSE for the local recall model.

t. We computed the root mean square error (RMSE) of the prediction of speed variations $\mathscr{E}_{\mathrm{ML}}(\tau)$, defined as

$$\mathscr{E}_{\mathrm{ML}}(\tau) = \sqrt{\frac{1}{n}\sum_{t=\Delta_s}^{n\Delta_s}(\mathbf{A}_{t,\mathrm{ML}} - \widehat{\mathbf{A}}_{t,\mathrm{ML}})^2}\ ,$$

We also computed $\tilde{\mathscr{R}}^2_{\mathrm{ML}}(\tau)$, the adjusted coefficient of determination (adjusted R2) of the model Richard (1994):

$$\mathscr{R}^2_{\mathrm{ML}}(\tau) = 1 - \frac{\sum_{t=\Delta_s}^{n\Delta_s}(\mathbf{A}_{t,\mathrm{ML}} - \widehat{\mathbf{A}}_{t,\mathrm{ML}})^2}{\sum_{t=\Delta_s}^{n\Delta_s}(\mathbf{A}_{t,\mathrm{ML}} - \overline{\mathbf{A}}_{\mathrm{ML}})^2}$$

$$\tilde{\mathscr{R}}^2_{\mathrm{ML}}(\tau) = 1 - (1 - \mathscr{R}^2_{\mathrm{ML}}(\tau))\frac{n-1}{n-1-p}\ .$$

where $\overline{\mathbf{A}}_{\mathrm{ML}}$ is the average value of the observed speed change of the CoP and p is the number of predictor variables in the model. $\tilde{\mathscr{R}}^2_{\mathrm{AP}}(\tau)$ and $\mathscr{E}_{\mathrm{AP}}(\tau)$ are computed similarly. All the aforementioned quantities are calculated at the trajectory level, and we analyzed the resulting distribution over

trajectories. Using these metrics, we compared the accuracy of the local recall model to two others, to highlight the benefits of the local recall approach. In the first one, called global recall, the CoP is assumed to follow a Langevin dynamic similar to (10.2.1), except that the recall force pulls the trajectory towards the center of the base of support:

$$d\mathscr{V}_t^{\mathrm{CoP}} = \left[\Lambda'(-\mathrm{CoP}_t) + \Gamma'(-\mathscr{V}_t^{\mathrm{CoP}})\right] dt + \Sigma' dB_t$$

In the second one, called average model, we assume that the CoP acceleration can be directly approximated by the CoM acceleration:

$$d\mathscr{V}_t^{\mathrm{CoP}} \approx d\mathscr{V}_t^{\mathrm{CoM}} + \Sigma'' dB_t$$

Note that Λ', Γ', Σ' and Σ'' were also estimated using the OLS algorithm.

Parameters distribution. In a second part, we compared the distributions of the estimated parameters Λ, Γ, Σ of the local recall model for different groups of individual (such as healthy young individuals and older adults), as well as for different balance conditions (open eyes and closed eyes), to show that the values of these parameters may be indicative of different postural control profiles.

10.3 Results

10.3.1 Model evaluation

In all our analyses, the local recall model produced significantly larger values of explained variance (see Figure 10.2). As shown in Figure 10.2(a), this improvement was observed for every recording condition of our experiment ($p < 10^{-8}$ compared to the global recall model, $p < 10^{-10}$ compared to the average model). Similar results were obtained on the larger population 2, which is the public dataset of dos Santos et al. (2017) (Figure 10.2(b), all respective p-values are $< 10^{-40}$). This is particularly interesting as the population included in this second dataset is larger and far more diverse, including young individuals and older adults, as well as individuals with a history of falls. It is also interesting to note that the average model, which tries to infer the acceleration of the CoP using solely the acceleration of the CoM, achieves adjusted coefficients of determination closed to zero, and significantly lower than the other models. This tends to show that while the CoP and CoM are closely related together, the CoP possesses its own dynamic that cannot be fully expressed by the CoM dynamic. Further analyses show that this improvement occurs for every trajectory (Figure 10.3, $p < 10^{-20}$). Similar improvements were observed for the RMSE metric (see Figure 10.4, $p < 10^{-20}$). This confirms that the local recall model provides better predictions of the CoP dynamics, for every recording condition (open/closed eyes, rigid surface and foam) and for every individual.

10.3.2 Estimated parameters distribution

Table 10.2 reports the average and standard deviation of the estimated values of the parameters of the local recall model – Λ, Γ and Σ – for different recording conditions and different groups of individuals. Interestingly, the perturbation coefficient Σ generally increases as the *expected* quality of the postural control decreases. For instance, AP perturbations for young individuals on rigid surface with open eyes $2.05(\pm 0.27)$ are significantly lower than the values for young individuals on foam surface with closed eyes $5.82(\pm 1.65)$ ($p < 10^{-20}$), which in turn are lower than the values for older adults on foam surface with closed eyes $9.51(\pm 3.59)$ ($p < 10^{-8}$). This relation is further explored in Figure 10.5, where it can be seen that this phenomenon is observable on both populations. Moreover, while no significant variations of Γ_{AP} are observed, Γ_{ML} shows an important decrease on

Pop. 1	Units	Open Eyes Rigid	Closed Eyes Rigid	Open Eyes Foam	Closed Eyes Foam
Recall ML	s^{-2}	35.17($\pm$11.54)	34.74($\pm$11.85)	35.08($\pm$11.01)	35.87($\pm$7.73)
Recall AP	s^{-2}	43.85($\pm$11.27)	49.00($\pm$13.71)	47.72($\pm$12.87)	49.70($\pm$6.48)
Damping ML	s^{-1}	7.49($\pm$2.16)	6.81($\pm$2.08)	6.07($\pm$2.44)	5.93($\pm$1.76)
Damping AP	s^{-1}	7.11($\pm$1.76)	6.67($\pm$1.94)	5.72($\pm$2.12)	5.77($\pm$0.91)
Perturbation ML	$(cm \times s^{-2})$	1.37($\pm$0.53)	1.39($\pm$0.32)	2.72($\pm$0.78)	4.23($\pm$1.69)
Perturbation AP	$(cm \times s^{-2})$	2.36($\pm$0.73)	2.82($\pm$0.83)	4.40($\pm$1.50)	7.32($\pm$3.14)
Pop 2. (YP)	Units	Open Eyes Rigid	Closed Eyes Rigid	Open Eyes Foam	Closed Eyes Foam
Recall ML	s^{-2}	40.32($\pm$8.11)	40.10($\pm$7.74)	37.91($\pm$6.25)	37.33($\pm$6.04)
Recall AP	s^{-2}	35.94($\pm$10.92)	39.86($\pm$11.42)	42.22($\pm$7.50)	48.85($\pm$9.21)
Damping ML	s^{-1}	9.74($\pm$3.73)	9.05($\pm$3.54)	4.49($\pm$1.10)	3.91($\pm$0.98)
Damping AP	s^{-1}	5.59($\pm$1.68)	5.03($\pm$1.46)	3.56($\pm$0.78)	3.96($\pm$1.08)
Perturbation ML	$cm \times s^{-2}$	1.96($\pm$0.27)	2.04($\pm$0.32)	3.06($\pm$0.45)	4.00($\pm$0.82)
Perturbation AP	$cm \times s^{-2}$	2.05($\pm$0.27)	2.27($\pm$0.40)	3.51($\pm$0.66)	5.82($\pm$1.65)
Pop 2. (OP)	Units	Open Eyes Rigid	Closed Eyes Rigid	Open Eyes Foam	Closed Eyes Foam
Recall ML	s^{-2}	35.21($\pm$6.90)	38.71($\pm$8.32)	31.53($\pm$5.40)	32.85($\pm$6.71)
Recall AP	s^{-2}	43.39($\pm$9.22)	47.26($\pm$11.79)	50.33($\pm$8.23)	55.67($\pm$12.47)
Damping ML	s^{-1}	8.20($\pm$3.39)	7.68($\pm$3.72)	3.45($\pm$0.94)	3.63($\pm$1.05)
Damping AP	s^{-1}	4.86($\pm$1.82)	4.75($\pm$1.86)	4.07($\pm$1.15)	4.43($\pm$1.31)
Perturbation ML	$cm \times s^{-2}$	2.03($\pm$0.33)	2.19($\pm$0.43)	3.54($\pm$0.83)	4.76($\pm$1.56)
Perturbation AP	$cm \times s^{-2}$	2.41($\pm$0.53)	2.93($\pm$0.70)	5.96($\pm$1.65)	9.51($\pm$3.59)

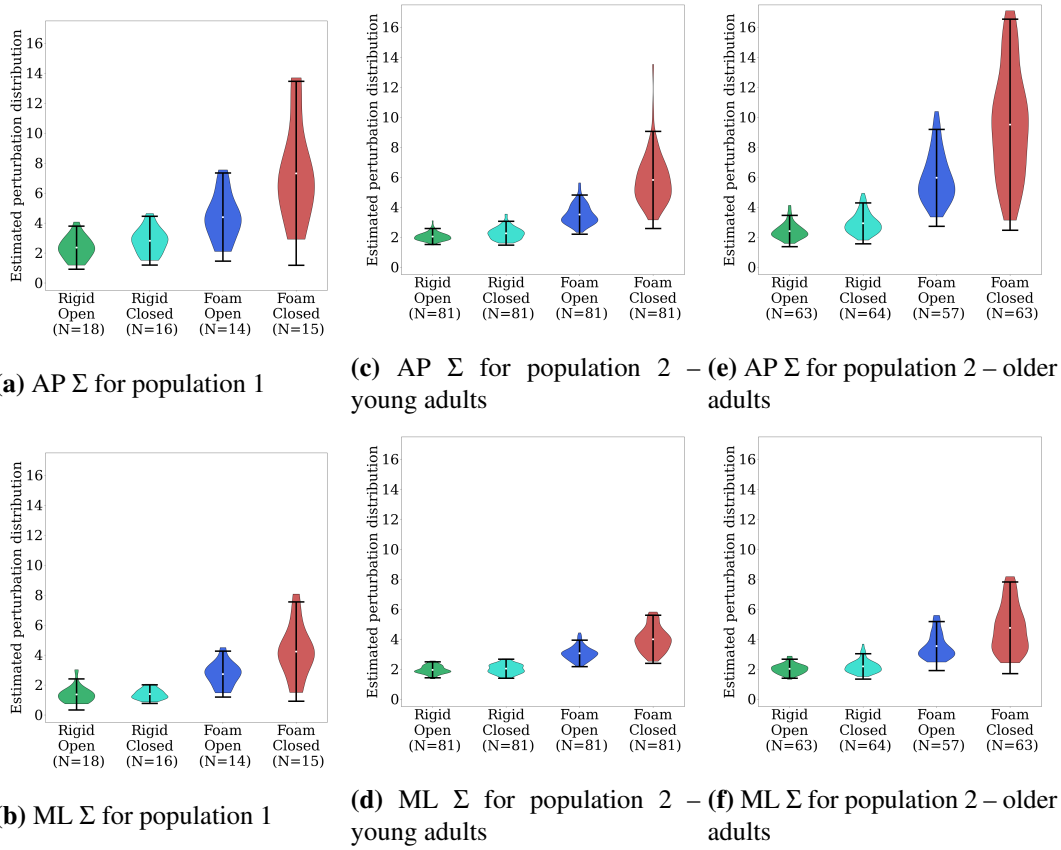
Table 10.2. Average ($\pm$ standard deviation) of the estimated parameters for the local recall model for different experimental conditions and for different populations: (top) Population 1, (middle for the young population (YP) and bottom for the (OP)) Population 2.

Pop 2. for individuals on foam surfaces compared to individuals on rigid surfaces ($p < 10^{-10}$). Note however that this difference is not observed on Pop 1 (see Figure 10.6). Conversely, the local recall coefficient Λ does not vary significantly between recording conditions. While a mild increase is observed on the AP axis between individuals on rigid surfaces and foam surfaces for Pop. 2 ($p < 0.01$), further analyses show strong overlap of the confidence intervals (see Figure 10.7). Therefore, these observations are insufficient to conclude in either direction.

10.4 Discussion

In the first part of the analyses we compared the accuracy of the predictions of three models : the global recall model, where the CoP follows a Langevin dynamic whose sole equilibrium point is the center of the force platform; the local recall model, where the CoP is assumed to follow a Langevin dynamic with the CoM position as non-static attachment point; and the average model, which assumes that the CoP acceleration is driven solely by the CoM acceleration.

The results of the analyses showed that the local recall model provided significantly better predictions of the CoP dynamic than its two counterparts. This is particularly true for the average model, whose accuracy is the lowest, which tends to show that the CoP acceleration cannot be approximated



(a) AP Σ for population 1

(c) AP Σ for population 2 – young adults

(e) AP Σ for population 2 – older adults

(b) ML Σ for population 1

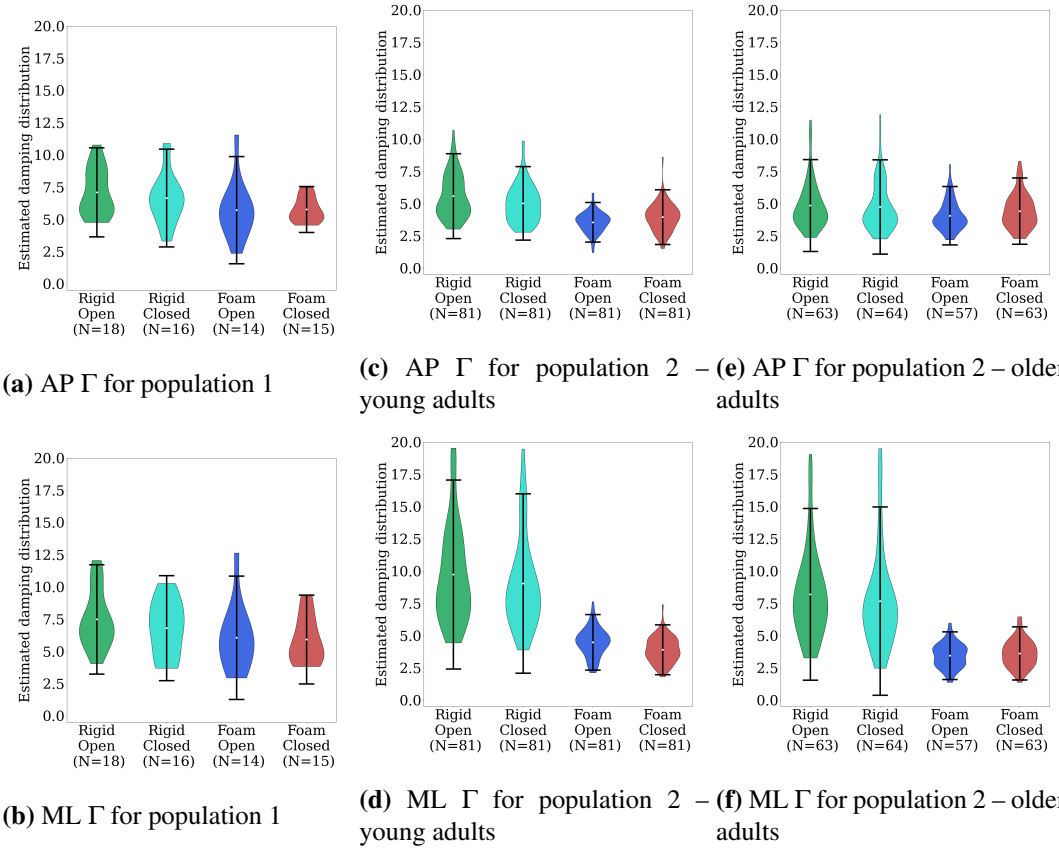
(d) ML Σ for population 2 – young adults

(f) ML Σ for population 2 – older adults

Figure 10.5. Distribution of the estimated perturbation coefficient in the local recall model for different conditions (green: open eyes rigid surface, cyan: closed eyes rigid surface, blue: open eyes foam and red: closed eyes foam), different populations (left: Population 1, middle: young individuals of population 2, right: older adults of population 2) and different axes (top: antero-posterior, bottom: medio-lateral. The whiskers indicate the 95% confidence interval. Average values of the Σ parameter significantly increase as the expected balance deteriorates.

by the CoM acceleration. However, the analyses also showed that by adding information about the CoM position in the Langevin model, the local recall model produces better estimates of the CoP dynamic than the global recall model. This result suggests that the trajectory of the CoM is important to understand the CoP dynamic, and that the Langevin model may provide relevant insights into the CoP behavior with respect to the CoM. Crucially, the predictions accuracy improvement was consistent, and occurred for each trajectory of both datasets, regardless of the protocol or of individual characteristics.

This observation that the CoM is key to understand the CoP dynamic is in line with the results discussed by previous works. In Zatsiorsky and Duarte (1999) a method was proposed to decompose the CoP trajectory in two components, rambling and trembling, where the latter is assumed to reflect the oscillations of the CoP around a reference point trajectory. In their findings, the authors mentioned that this reference trajectory – computed as the interpolation of the CoP points at which the horizontal force resultant vanishes – is very close to the CoM trajectory, and this result can be seen as hinting at the possibility that the CoM might be a reference point around which the CoP gravitates. In Winter et al. (1998) the authors introduced a model based on the inverted pendulum that relies on the assumption that muscles of the ankle act as springs to cause the CoP to control the body deviations from the vertical. This hypothesis has been successfully applied to continuous



(a) AP Γ for population 1

(c) AP Γ for population 2 – young adults

(e) AP Γ for population 2 – older adults

(b) ML Γ for population 1

(d) ML Γ for population 2 – young adults

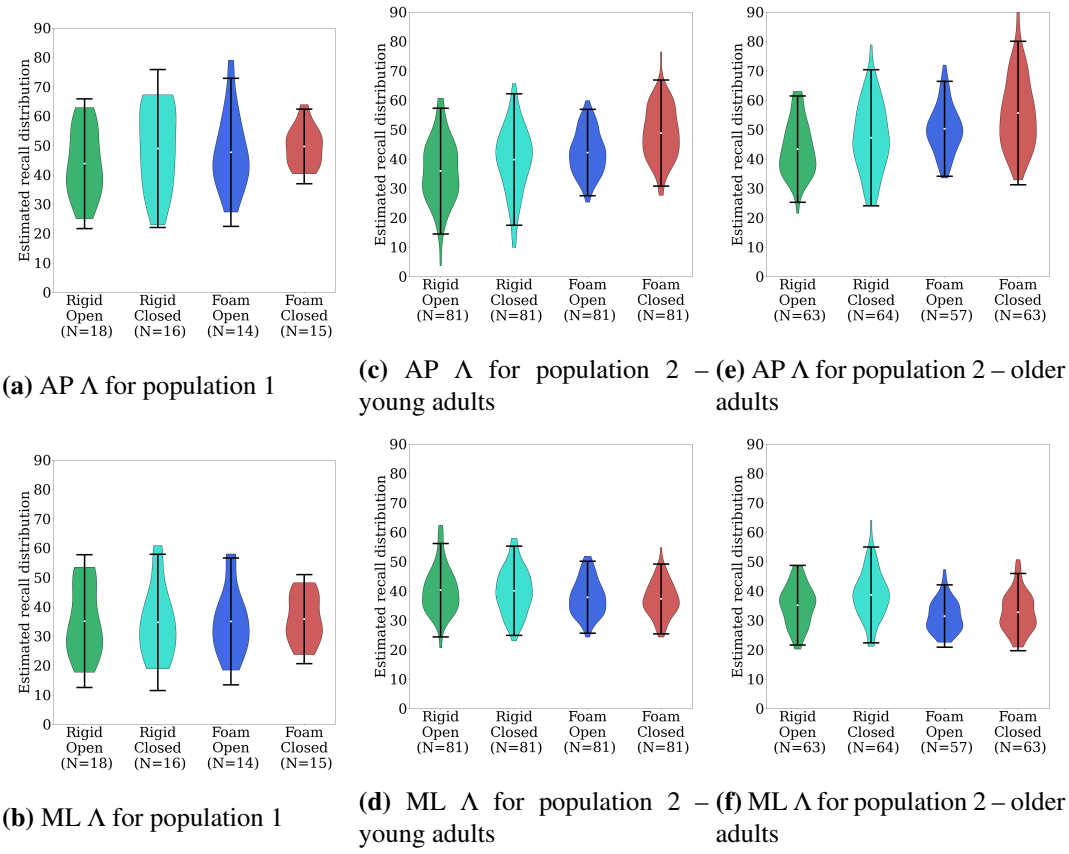
(f) ML Γ for population 2 – older adults

Figure 10.6. Distribution of the estimated damping coefficient in the local recall model for different conditions (green: open eyes rigid surface, cyan: closed eyes rigid surface, blue: open eyes foam and red: closed eyes foam), different populations (left: Population 1, middle: young individuals of population 2, right: older adults of population 2) and different axes (top: antero-posterior, bottom: medio-lateral. The whiskers indicate the 95% confidence interval. While no significant variations are observed in Population 1, Γ_{ML} significantly decreases between the rigid and foam conditions in Population 2.

linear feedback controllers such as PID (Proportional, Integral, Derivative) systems Mahboobin et al. (2007); Masani et al. (2006); Peterka (2002). In these works the CoM is central to the model, as the forces to maintain posture are modeled by springs dependent on both body angle and body angular velocity, which can be assumed approximately proportional to the CoM position and speed for small body angles.

In our model the mediolateral and anteroposterior components are assumed to have distinct dynamics. This assumption is important considering the significant differences in balance control observed in each axis through the characteristics of the CoP Baratto et al. (2002) or the different muscles involved Winter et al. (1998). However, we additionally assume that those dynamics are independent. While this assumption is a commonly used approximation Collins and De Luca (1993), previous works Bosek et al. (2004) have proposed models where the dynamic of the postural control is influenced by the radius, i.e. the distance between the CoP and the center of the base of support. As the radius inherently depends simultaneously on both the AP and ML coordinates, such phenomenon cannot be captured by the local recall model, and studying extensions of the model that can embed possible axes interdependence seems an interesting research direction.

It should be noted that RMSE and R2, while providing important information, are not perfect metrics to evaluate the accuracy of one model. Indeed, the problem of measuring goodness-of-fit is



(a) AP Λ for population 1

(c) AP Λ for population 2 – young adults

(e) AP Λ for population 2 – older adults

(b) ML Λ for population 1

(d) ML Λ for population 2 – young adults

(f) ML Λ for population 2 – older adults

Figure 10.7. Distribution of the estimated recall coefficient in the local recall model for different conditions (green: open eyes rigid surface, cyan: closed eyes rigid surface, blue: open eyes foam and red: closed eyes foam), different populations (left: Population 1, middle: young individuals of population 2, right: older adults of population 2) and different axes (top: antero-posterior, bottom: medio-lateral. The whiskers indicate the 95% confidence interval. No significant variations of Λ are observed in either populations.

still an open research topic (see e.g. Jitkrittum et al. (2017)). Consequently, while observed results tend to show that the local recall model is better than its two counterparts, i.e. at a relative scale, it is significantly harder to assess how good the model is on an absolute scale. For instance, model residuals are ambivalent as they encompass both model errors, i.e. inaccurate predictions, as well as the Brownian perturbations that can be part of the postural control system. Previous works have considered alternative approaches to validate models, such as bootstrapping. This method uses a generative model to assess the likelihood of the observed characteristics on the original signals compared to the characteristics of the generated ones. Recently, this method of validation has been applied on a Langevin equation of the CoP Tawaki and Murakami (2019). Unfortunately, this approach requires the choice of specific characteristics of the CoP, a choice which can have a significant influence on the results, and also necessitates a joint generative model of both the CoP and the CoM. Nevertheless, this is an interesting future direction for this research.

Interestingly, the parameter Λ, which encodes the strength of the recall force in the local recall model, does not significantly vary between different groups and protocols in our experiments. In previous works, the recall parameter has been interpreted as related to the ankle joint stiffness Hernandez et al. (2015); Lauk et al. (1999), which is defined as the derivative of the torque applied at the ankle with respect to the angle of deviation from the gravity line Winter et al. (2001). However, Langevin models do not explicitly incorporate ankle stiffness and further work is required to

link quantitatively the recall parameter to biomechanics components.

Conversely, the parameter Γ, which measures the strength of the damping, i.e. the force which opposes to the velocity of the system, was shown in our experiments on Pop 2. to decrease in the ML axis when going from a rigid surface to a foam surface. This result shows that this force, preventing the velocity of the body becoming too strong, is a component of postural control that can be vulnerable to sensorial perturbations induced by unstable surfaces Kiers et al. (2012), and thus be indicative of some specificity in sensorimotor profile. The fact that this result is not observed on Pop 1. could be explained by some differences existing between the type of foam used in each dataset.

Finally, the parameter Σ appears to strongly increase when the expected balance quality decreases in our experiments. The interpretation of the perturbation force is more complicated, as it can be representative of two distinct phenomena. On the one hand, Σ corresponds to the coefficient of diffusion associated with the Brownian process ΣB_t and therefore may be associated with the strength of the stochastic activity in the postural system, which has been suggested to increase with aging Collins et al. (1995). These perturbations may arise from various sources such as breathing Bottaro et al. (2005); Hodges et al. (2002) or from any errors in sensorimotor integration or postural adjustments. On the other hand, ΣdB_t also represents the noise in the formalization of the linear model, and therefore can include the fitting error of the model. Consequently the parameter Σ may also reflect a wrong adjustment to the local recall model. In both views, large components of Σ for a trajectory in comparison to other individual's trajectories could be the sign of a bad balance, either because of a perturbed postural control, or because the individual does not share the same postural control dynamics as others, which could be explained by the existence of age-dependent postural control strategies Collins et al. (1995).

10.5 Conclusion

In conclusion, this study showed that the dynamic of the CoP is strongly influenced by the trajectory of the CoM, and that the spring restorative force that is part of Langevin models for quiet stance should be aimed toward to the CoM position, instead of the center of the base of support. We showed that this modified model, called local recall, significantly increased the accuracy of the prediction of the CoP Langevin model, providing interesting research directions for future postural control models. Additionally, we provided a method to estimate the parameters of the local recall model, and showed that key parameters (damping, perturbations) are closely correlated with the quality of postural control. Finally, these findings support the hypothesis that the Langevin model has the potential to quantify interesting aspects of postural control, and could be significantly improved by the embedding of the CoM trajectory. Further work is needed to link precisely the parameters of the local recall model to biomechanical components of postural control as well as to investigate if other improvements of the model, such as the addition of nonlinearities Gottschall et al. (2009) or intermittent postural adjustements Bottaro et al. (2005), could lead to a more realistic quantification of postural control.

Conclusion and Perspectives

This work presents a new multivariate and nonparametric generalization of rank statistics through learning-to-rank approaches. An in-depth theoretical investigation led to a series of statistical learning guarantees in the general formulation and for the two-sample problem. Also, we presented how this generic framework encompasses interesting interpretability by leveraging the ROC analysis. Lastly, various (modeling) applications were presented in machine learning and biomedicine with associated algorithmic procedures. In particular, two contributions in the context of the Human postural analysis are gathered: a method for the statistical comparison between clinical populations and a stochastic model to generate posturographic data. The main results are sequentially summarized, and selected perspectives are detailed to conclude this manuscript. In particular, potential applications and theoretical open questions are highlighted that remain not studied yet to the best of our knowledge.

General conclusion. The first part of this manuscript reviews problems with two samples and the necessary theoretical tools to derive the results related to two-sample linear R-processes. Precisely, Chapter 2 gathers classic paradigms from both statistics and machine learning, that are later shown to be applications of our proposed class of R-processes. While one can find all results and approaches in the indicated references with greater detail, we promote a unified formulation that aims to help for a better understanding and comparison of the state-of-the-art methods. In particular, we raise their limits and the necessity for a new methodology, such that classic statistical properties hold true under mild probabilistic conditions while allowing for (simple) practical implementations.

In Chapter 3, the theoretical tools for such problems are recalled, particularly by motivating them in the ERM context. Classic ERM approaches study statistical risk criteria taking the form of empirical processes analyzed using PAC nonasymptotic techniques. The risk is formulated as M-statistics. However, risk criteria of higher order can be decomposed using U-statistics. We gather some results on one and two-sample U-processes, reformulated under the assumptions of this manuscript that are key for the following chapters.

The second part is the core of this work. We introduced and constructed a new generic framework for two-sample linear rank statistics. A multivariate and nonparametric generalization of the univariate statistic is obtained using bipartite ranking approaches. This generalization is proved to be at the crossroads of hypothesis testing, learning-to-rank methods, and ROC analysis. Additionally, it allows for great adaptability in many applications while inheriting from the univariate theoretical guarantees.

Chapter 5 details and establishes uniform generalization bounds and related results for these R-processes. We first presented the optimal elements that maximize the statistic. Then, the uniform deviations of the statistic *w.r.t.* its continuous counterpart, referred to as W_ϕ-ranking performance criterion, are proved to achieve a classic learning rate bound of order $O_{\mathbb{P}}(1/\sqrt{N})$, when based on a training dataset of size N. Additionally, similar contributions for model selection procedures and smoothed versions of the statistics are detailed. These results are new in the statistical learning literature and provide distribution-free nonasymptotic bounds. The constants are detailed and only depend on the parameters of the statistic and the 'theoretical' proportion of one sample among the pooled samples. Importantly, this formulation is one of the first grounded multivariate generalizations of R-statistics for nonparametric probabilistic assumptions.

All these results rely on the uniform concentration bound for two-sample U-processes proved in Chapter 4. It is a new step towards the nonasymptotic analysis of uniform deviations of U-statistics based on multiple samples drawn from different distributions when indexed by infinite classes of kernels. For these two chapters, we used a series of techniques for the proofs, ranging from stochastic processes such as U-processes, to chaining and decoupling methods for non-*i.i.d.* statistics, under

minimal assumptions on the underlying probability distributions of the two samples, on the class $\mathscr{S}$ and on the function ϕ.

Chapter 6 states a new generic framework for the two-sample problem. A two-stage in proposed wherein, given two statistical samples: (1) the optimal scoring function on the first halves of the samples are learnt by a bipartite ranking algorithm, (2) the corresponding R-statistic is used to perform the statistical test by mapping the second ones onto the real line. While the first step has strong guarantees obtained in Chapter 5, we proved nonasymptotic control of the bias for the second step. Also the asymptotic laws of the statistic under both testing hypotheses are proved. Our class of R-statistics inherits from the strong properties of the univariate formulation while allowing for high-dimensional and generic feature spaces (*i.e.* not necessarily Euclidean). Additionally, it encompasses a series of classic homogeneity testing frameworks and provides for multiple algorithmic architectures. These results provide a competitive approach for the two-sample problem, compared to state-of-the-art methods as detailed in Chapter 2. This procedure is distribution-free under the null, with bounds independent on the dimension of the feature space under both hypotheses while achieving exact control of the type-I and type-II errors. In particular, it inherits from the robustness and unbiasedness of univariate ranks statistics.

In Chapter 7, we explored algorithmic empirical properties of the proposed procedures for bipartite ranking and the two-sample problem. First, formulated as a scalar performance criterion for the bipartite ranking problem, we explored the importance of the score-generating function ϕ for the statistic. In particular, we implemented an exact optimization gradient ascent algorithm. Then, we implemented and tested a series of algorithmic possibilities for the high-dimensional two-sample problem *w.r.t.* the choice of state-of-the-art bipartite ranking algorithms. We compared the results to three classic tests that rely on different approaches from ours. All codes are open source and available online, promoting the replicability of the experiments.

The last part gathers three applications that highlight the adaptivity of the proposed R-processes. Chapter 8 explores anomaly ranking modeling, wherein our framework allows for a generalization of scalar performance rank-based criteria that are intrinsic to the model, compared to state-of-the-art methods. It bridges the gap from unsupervised one-sample to two-sample frameworks.

Chapter 9 provides a new algorithmic methodology for the two-sample problem when considering biomedical data samples. In particular, we were interested in addressing this problem with an interpretable algorithmic procedure. We compared our results to traditional methods used in the biomedical community when applied to real statokinesigrams. It reveals an essential difference in the statistical findings while being consistent with those from a typical multivariate two-sample test.

Lastly, regarding Chapter 10, a new stochastic generative model for statokinesigrams is proposed. Based on Langevin stochastic differential equations, it models and couples the temporal evolution of two indicators: the Center of Pressure and the Center of Mass. Importantly, we show how this approach can be used as a generative model, while providing an estimation of the parameters proper to the clinical cohort or the patients. Hence, it is a step forward for the modeling and understanding of such stochastic dynamics.

Perspectives and open questions. This last paragraph outlines selected research directions our results lead to. In particular, we develop ongoing topics of research related to the two-sample problem and to the question of *choosing the optimal* test statistic.

- *Chapter 5.* The series of results are obtained under particular assumptions on the type of classes of scoring functions ($\mathscr{S}_0$), characterized by Vapnik-Chervonenkis theoretical tools. However, typical examples used in machine learning algorithms do not satisfy these requisites.

A possible extension is related to using Rademacher risk criteria, allowing localization bounds for instance. We want to explore this very interesting direction despite the high technical complexity.

- *Chapter 6*. From the practical perspective, the proposed procedure offers great algorithmic possibilities depending on the field of applications or the type of data. The theoretical guarantees are proved for generic forms of: the bipartite ranking algorithm ($\mathscr{A}$), the score-generating functions (ϕ), and the class of scoring function ($\mathscr{S}_0$). For instance, Chapter 9 adapted it to a bagging-based two-sample test for biomedical data.

- *Chapter 7*. From a more general perspective, we discovered a lack of practical methods for pairwise ranking algorithms in the machine learning community. We detailed some research directions in Section 2.2.2, Chap. 2. However, they correspond to the particular choices of $\phi(u) = u$, and when $\mathscr{S}_0$ is composed of linear functions. We highlight the intrinsic computational difficulty, as it has to be balanced with high algorithmic complexity due to the pairs comparisons. This point falls into the lack of bivariate algorithms, also noticed for U-statistics optimization in recent publications for instance.

- *Chapters 9 and 10*. Lastly, ongoing research directions are related to postural control analysis, either through statistical methods for modeling/estimating it or *via* the ability to propose a generative model. We work on pairwise statistical models applied to follow-ups, *i.e.*, longitudinal data, to compare trajectories of patients. Also, we have promising results on implementing the proposed deterministic algorithm to maximize the W_ϕ criterion to biomedical model validation. We believe in the importance of interpretable and replicable algorithmic procedures, seeking more reliable results.

This manuscript gathers research works and related results to two-sample linear R-statistics wherein we considered fixed score-generating functions ϕ. However, and motivated by many cited articles, *e.g.*, Clémençon and Vayatis (2007); Rudin (2006), we detailed how the choice of ϕ is fundamental to reveal various characteristics of the studied model throughout the manuscript. Also, from the univariate R-statistics literature, typical examples of *optimal* functions ϕ exist for maximizing asymptotic efficiencies depending on the problem, see *e.g.* examples of Chapter 13 and Section 15.5 in van der Vaart (1998). The following paragraph describes selected research directions on choosing the *best* score-generating function. We accompany the following points with preliminary results and applications regrouped in Appendix chapter A.

Which is the best R-statistic we can choose?

- *Theoretical perspective.* The works of H. Koul (see Koul (2002)), study asymptotic uniform deviations of linear rank statistics on the class of score-generating functions. However, his studies focus on the particular multivariate regression model. In the continuity of our framework, we believe that similar nonasymptotic results are important and can be obtained. More precisely, we aim to derive nonasymptotic generalization bounds of the deviations of the generic R-statistic when indexed by a class of score-generating functions. This would help to better understand the empirical W_ϕ-criteria and to obtain refined results on its stochastic fluctuations. In particular, we hope to express explicit nonasymptotic bounds depending on the same parameters as the ones of Chap. 5: p, $\mathscr{S}_0$, ϕ. This work requires broader and more complex techniques than the ones of Chap. 5. Some preliminary results are detailed in Appendix section A.1.

- *Adaptive two-sample homogeneity rank tests.* We highlighted the adaptivity of the two-sample procedure proposed in Chap. 6, for the choice of ϕ in points 3 (for the homogeneity test statistic) and 4 (for the bipartite ranking). Motivated by the empirical results to maximize the empirical W_ϕ-criteria of Chap. 7, we propose an alternative approach to the two-stage procedure in Appendix section A.2. Instead of using a bipartite ranking algorithm, we propose to maximize the exact W_ϕ-criterion at *Step 1*. We also describe an adaptive test statistic that aggregates a sequence of R-statistics indexed by a class of functions ϕ. It aims at choosing the *best* R-statistic depending on ϕ, to maximize the power at the least conservative level of test bounded by α. In particular, we would like to derive theoretical guarantees on this type of procedures.

- *Towards an efficient optimization algorithm for high-dimensional two-sample testing.* We discussed at length about the lack of methods for the multivariate two-sample problem for complex data structures. In particular, and motivated by biomedical applications, typical examples encompass very high-dimensional settings (*e.g.*, when the feature dimension and sample size are of similar order ($n \sim d$) if $\mathbb{R}^d$) and small sample sizes. For such frameworks, obtaining replicable and interpretable results is even more difficult. In this sense, we gather in the Appendix section A.3 first numerical results. We sampled two samples of sizes $n = m = 30$ in $\mathbb{R}^d$ with $d \in \{30, 50\}$, such that we can perform the exact maximization of W_ϕ (*Step 1*) for various choices of ϕ using the gradient ascent algorithm of Chap. 5 on subsamples of size $n = m = 24$. The remaining observations are used for the homogeneity testing with the related R-statistic. Therefore, we can easily compute the exact null distribution of the statistics for such samples sizes. This is an ongoing research direction as it would allow for the exact generalization of R-statistics. Importantly, this procedure has tractable optimization algorithm and exact homogeneity test statistics that allows for very high-dimensional data analysis.

Bibliography

Important facts about falls. Technical report, Centers for Disease Control and Prevention, 2017.

R. Adamczak. Moment inequalities for U-statistics. *The Annals of Probability*, 34(6):2288 – 2314, 2006.

S. Agarwal. *The Infinite Push: A New Support Vector Ranking Algorithm that Directly Optimizes Accuracy at the Absolute Top of the List*, pages 839–850. 2011.

S. Agarwal. Surrogate regret bounds for bipartite ranking via strongly proper losses. *Journal of Machine Learning Research*, 15(49):1653–1674, 2014.

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

R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. *SIGMOD Rec.*, 27(2):94–105, 1998.

N. Ailon and M. Mohri. An efficient reduction of ranking to classification. In *21st Annual Conference on Learning Theory - COLT*, pages 87–98, 2008.

M. Albert, B. Laurent, A. Marrel, and A. Meynaoui. Adaptive test of independence based on hsic measures, 2021.

N. B. Alexander, N. Shepard, M. J. Gu, and A. Schultz. Postural control in young and elderly adults when stance is perturbed: kinematics. *Journal of Gerontology*, 47(3):M79–M87, 1992.

O. Alghushairy, R. Alsini, T. Soule, and X. Ma. A review of local outlier factor algorithms for outlier detection in big data streams. *Big Data and Cognitive Computing*, 5(1), 2021.

M. A. Arcones and E. Giné. U-processes indexed by vapnik-cervonenkis classes of functions with applications to asymptotics and bootstrap of *u*-statistics with estimated parameters. *Stochastic Processes and their Applications*, 52(1):17–38, 1994. ISSN 0304-4149.

E. Arias-Castro and M. Wang. Distribution-free tests for sparse heterogeneous mixtures, 2017.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

J. Audiffren and E. Contal. Preprocessing the nintendo wii board signal to derive more accurate descriptors of statokinesigrams. *Sensors*, 16(8), 2016.

J. Audiffren, I. Bargiotas, N. Vayatis, P. Vidal, and D. Ricard. A non linear scoring approach for evaluating balance: classification of elderly as fallers and non-fallers. *Plos One*, 11:12, 2016.

F. Bach, Z. Harchaoui, and E. Moulines. Testing for homogeneity with kernel Fischer discriminant analysis. In *Advances in Neural Information Processing Systems 20*. MIT Press, Cambridge, MA, 2008.

M. Baker and D. Penny. Is there a reproducibility crisis? *Nature*, 533(7604):452–454, 2016. ISSN 14764687.

M.-F. Balcan, N. Bansal, A. Beygelzimer, D. Coppersmith, J. Langford, and J. B. Sorkin. Robust reductions from ranking to classification. In *Learning Theory*, pages 604–619. Springer Berlin Heidelberg, 2007.

L. Baratto, P. G. Morasso, C. Re, and G. Spada. A new look at posturographic analysis in the clinical context: sway-density versus other parameterization techniques. *Motor control*, 6(3):246–270, 2002.

Y. Baraud. Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8(5):577 – 606, 2002.

I. Bargiotas, J. Audiffren, N. Vayatis, P. Vidal, S. Buffat, and A. Yelnik. On the importance of local dynamics in statokinesigram: A multivariate approach for postural control evaluation in elderly. *Plos One*, 13(2), 2018.

I. Bargiotas, A. Moreau, A. Vienne, F. Bompaire, M. Baruteau, and M. de Laage. Balance impairment in radiation induced leukoencephalopathy patients is coupled with altered visual attention in natural tasks. *Frontiers in Neurology*, 9:1185, 2019.

I. Bargiotas, A. Kalogeratos, M. Limnios, P.-P. Vidal, D. Ricard, and N. Vayatis. Revealing posturographic profile of patients with Parkinsonian syndromes through a novel hypothesis testing framework based on machine learning. *PLOS ONE*, 16(2):1–22, 02 2021.

L. Baringhaus and C. Franz. On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88(1):190–206, 2004.

V. Barnett and T. Lewis. *Outliers in Statistical Data. 3rd edition*. Probability and Mathematical Statistics. 1994.

P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2003.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

D. Basu. The family of ancillary statistics. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 21(3/4):247–256, 1959.

C. G. Begley and L. M. Ellis. Drug development: Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533, 2012. ISSN 00280836.

J. Beirlant, S. Buitendag, E. del Barrio, M. Hallin, and F. Kamper. Center-outward quantiles and the measurement of multivariate risk. *Insurance: Mathematics and Economics*, 95:79–100, 2020. ISSN 0167-6687.

D. J. Benjamin, J. O. Berger, M. Johannesson, B. A. Nosek, E. J. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, C. Camerer, D. Cesarini, C. D. Chambers, M. Clyde, T. D. Cook, P. De Boeck, Z. Dienes, A. Dreber, K. Easwaran, C. Efferson, E. Fehr, F. Fidler, A. P. Field, M. Forster, E. I. George, R. Gonzalez, S. Goodman, E. Green, D. P. Green, A. G. Greenwald, J. D. Hadfield, L. V. Hedges, L. Held, T. Hua Ho, H. Hoijtink, D. J. Hruschka, K. Imai, G. Imbens, J. P. Ioannidis, M. Jeon, J. H. Jones, M. Kirchler, D. Laibson, J. List, R. Little, A. Lupia, E. Machery, S. E. Maxwell, M. McCarthy, D. A. Moore, S. L. Morgan, M. Munafó, S. Nakagawa, B. Nyhan, T. H. Parker, L. Pericchi, M. Perugini, J. Rouder, J. Rousseau, V. Savalei, F. D. Schönbrodt, T. Sellke, B. Sinclair, D. Tingley, T. Van Zandt, S. Vazire, D. J. Watts, C. Winship, R. L. Wolpert, Y. Xie, C. Young, J. Zinman, and V. E. Johnson. Redefine statistical significance. *Nature Human Behaviour*, 2(1): 6–10, 2018.

Y. Benjamini, R. D. D. Veaux, B. Efron, S. Evans, M. Glickman, B. I. Graubard, X. He, X.-L. Meng, N. Reid, S. M. Stigler, S. B. Vardeman, C. K. Wikle, T. Wright, L. J. Young, and K. Kafadar. The ASA president's task force statement on statistical significance and replicability. *The Annals of Applied Statistics*, 15(3):1084 – 1085, 2021.

G. Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962. ISSN 01621459.

L. Bergman and Y. Hoshen. Classification-Based Anomaly Detection for General Data. *arXiv:2005.02359*, May 2020.

A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer, Boston, MA, 2004.

S. N. Bernstein. *The Theory of Probabilities*. Gastehizdat Publishing House, 1946.

T. B. Berrett, I. Kontoyiannis, and R. J. Samworth. Optimal rates for independence testing via U-statistic permutation tests. *The Annals of Statistics*, 49(5):2457 – 2490, 2021. doi: 10.1214/20-AOS2041.

P. Bertail, S. Clémençon, Y. Guyonvarch, and N. Noiry. Learning from biased data: A semi-parametric approach. In *ICML*, 2021.

B. B. Bhattacharya. A general asymptotic framework for distribution-free graph-based two-sample tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(3):575–602, 2019.

G. Bhattacharya, K. Ghosh, and A. S. Chowdhury. Outlier detection using neighborhood rank difference. *Pattern Recognition Letters*, 60:24–31, 2015.

G. Biau and L. Gyorfi. On the asymptotic properties of a nonparametric l_1-test statistic of homogeneity. *IEEE Transactions on Information Theory*, 51(11):3965–3973, 2005.

P. J. Bickel. A Distribution Free Version of the Smirnov Two Sample Test in the p-Variate Case. *The Annals of Mathematical Statistics*, 40(1):1 – 23, 1969.

L. Birgé and P. Massart. Minimum contrast estimators on sieves: Exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.

J. M. Bland and D. G. Altman. Misleading statistics: Errors in textbooks, software and manuals. *International Journal of Epidemiology*, 17(2):245–247, 1988. ISSN 0300-5771.

J. Błaszczyk, R. Orawiec, D. Duda-Kłodowska, and G. Opala. Assessment of postural instability in patients with parkinson's disease. *Experimental Brain Research*, 183(1):107–114, 2007.

M. Bosek, B. Grzegorzewski, and A. Kowalczyk. Two-dimensional langevin approach to the human stabilogram. *Human movement science*, 22(6):649–660, 2004.

M. Bosek, B. Grzegorzewski, A. Kowalczyk, and I. Lubiński. Degradation of postural control system as a consequence of parkinson's disease and ageing. *Neuroscience letters*, 376(3):215–220, 2005.

A. Bottaro, M. Casadio, P. G. Morasso, and V. Sanguineti. Body sway during quiet standing: Is it the residual chattering of an intermittent stabilization process? *Human movement science*, 24(4): 588–615, 2005.

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of Classification: A Survey of Some Recent Advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

S. Boyd, C. Cortes, M. Mohri, and A. Radovanovic. Accuracy at the top. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

L. Breiman. Out-of-bag estimation. *Technical report*, 1996.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

M. Breunig, H. Kriegel, R. Ng, and J. Sander. Lof: identifying density-based local outliers. In *ACM sigmod record*, volume 29, pages 93–104, 2000.

T. Brus. A recurrence formula for the distribution of the wilcoxon rank sum statistic. *Statistics and Probability Letters*, 7(2):161–165, 1988. ISSN 0167-7152.

C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, page 89–96, 2005.

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

L. Cam. *Locally Asymptotically Normal Families of Distributions: Certain Approximations to Families of Distributions and Their Use in the Theory of Estimation and Testing Hypotheses*. University of California Berkeley, Calif: University of California publications in statistics. University of California Press, 1960.

O. Caron, T. Gélat, P. Rougier, and J.-P. Blanchi. A comparative analysis of the center of gravity and center of pressure trajectory path lengths in standing posture: an estimation of active stiffness. *Journal of applied biomechanics*, 16(3):234–247, 2000.

A. Carpentier, O. Collier, L. Comminges, A. Tsybakov, and Y. Wang. Minimax rate of testing in sparse linear regression. *Automation and Remote Control*, 80, 2018.

A. Carrington, P. Fieguth, H. Qazi, A. Holzinger, H. Chen, F. Mayr, and D. Manuel. A new concordant partial auc and partial c statistic for imbalanced data in the evaluation of machine learning algorithms. *BMC Med Inform Decis Mak*, 20(1), 2020.

V. R. Carvalho, J. L. Elsas, W. W. Cohen, and J. G. Carbonell. Suppressing outliers in pairwise preference ranking. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, page 1487–1488. Association for Computing Machinery, 2008.

L. Castelli, L. Stocchi, M. Patrignani, G. Sellitto, M. Giuliani, and L. Prosperini. We-measure: Toward a low-cost portable posturography for patients with multiple sclerosis using the commercial wii balance board. *Journal of the Neurological Sciences*, 359:440–444, 2015.

J. Chagdes, S. Rietdyk, J. Haddad, H. Zelaznik, A. Raman, K. Rhea, and T. A. Silver. Multiple timescales in postural dynamics associated with vision and a secondary task are revealed by wavelet analysis. *Experimental Brain Research*, 197(3):297–310, 2009.

A. Chakraborty and P. Chaudhuri. Tests for high-dimensional data based on means, spatial signs and spatial ranks. *The Annals of Statistics*, 45(2):771 – 799, 2017.

D. K. Chang. A note on the distribution of the wilcoxon rank sum statistic. *Statistics and Probability Letters*, 13(5):343–349, 1992.

P. Chaudhuri. On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association*, 91(434):862–872, 1996.

B. Chen, P. Liu, F. Xiao, Z. Liu, and Y. Wang. Review of the upright balance assessment based on the force plate. *International Journal of Environmental Research and Public Health*, 18(5), 2021.

S. X. Chen and Y.-L. Qin. A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics*, 38(2):808 – 835, 2010.

H. Chernoff. A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the sum of Observations. *The Annals of Mathematical Statistics*, 23(4):493 – 507, 1952.

H. Chernoff and I. Savage. Asymptotic normality and efficiency of certain non parametric test statistics. *Ann. Math. Stat.*, 29:972–994, 1958.

V. Chernozhukov, A. Galichon, M. Hallin, and M. Henry. Monge–kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1):223–256, 2017.

Y. Cheung and J. Klotz. The Mann Whitney Wilcoxon distribution using linked list. *Statistica Sinica*, 7:805–813, 1997.

E.-Y. Chung and J. P. Romano. Exact and asymptotically robust permutation tests. *The Annals of Statistics*, 41(2):484 – 507, 2013.

E.-Y. Chung and J. P. Romano. Asymptotically valid and exact permutation tests based on two-sample u-statistics. *Journal of Statistical Planning and Inference*, 168:97–105, 2016. ISSN 0378-3758.

K. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton. Fast two-sample testing with analytic representations of probability measures. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, page 1981–1989, 2015.

K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit, 2016.

R. Clark, A. Bryant, Y. Pua, P. McCrory, K. Bennell, and M. Hunt. Validity and reliability of the nintendo wii balance board for assessment of standing balance. *Gait and posture*, 31(3):307–310, 2010.

S. Clémençon. A statistical view of clustering performance through the theory of U-processes. *Journal of Multivariate Analysis*, 124:42–56, 2014.

S. Clémençon and N.Vayatis. Adaptive estimation of the optimal roc curve and a bipartite ranking algorithm. In *Proceedings of ALT'09*, 2009.

S. Clémençon and S. Robbiano. Minimax learning rates for bipartite ranking and plug-in rules. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 441–448, 2011.

S. Clémençon and S. Robbiano. Anomaly ranking as supervised bipartite ranking. In *Proceedings of the 31st International Conference on International Conference on Machine Learning*, volume 32 of *ICML'14*, page II–343–II–351, 2014.

S. Clémençon and S. Robbiano. The TreeRank Tournament Algorithm for Multipartite Ranking. *Journal of Nonparametric Statistics*, 27(1):107–126, 2015.

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

S. Clémençon and N. Vayatis. Tree-structured ranking rules and approximation of the optimal ROC curve. In *ALT '08: Proceedings of the 2008 conference on Algorithmic Learning Theory*, 2008.

S. Clémençon and N. Vayatis. Empirical performance maximization based on linear rank statistics. In *Advances in Neural Information Processing Systems*, volume 3559 of *Lecture Notes in Computer Science*, pages 1–15. Springer, 2009a.

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

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

S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and scoring using empirical risk minimization. In P. Auer and R. Meir, editors, *Proceedings of COLT 2005*, volume 3559 of *Lecture Notes in Computer Science*, pages 1–15. Springer, 2005.

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

S. Clémençon, M. Depecker, and N. Vayatis. AUC maximization and the two-sample problem. In *Advances in Neural Information Processing Systems*, volume 3559 of *Lecture Notes in Computer Science*, pages 1–15. Springer, 2009.

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

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

S. Clémençon, S. Robbiano, and N. Vayatis. Ranking Data with Ordinal Labels: Optimality and Pairwise Aggregation. *Machine Learning*, 93(1):67–104, 2013b.

S. Clémencon and N. Vayatis. On partitioning rules for bipartite ranking. In *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 97–104. PMLR, 2009.

S. Clémençon and J. Jakubowicz. Scoring anomalies: a M-estimation formulation. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 659–667. Proceedings of Machine Learning Research, 2013.

S. Clémençon and A. Thomas. Mass volume curves and anomaly ranking. *Electronic Journal of Statistics*, 12(2):2806 – 2872, 2018.

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

J. Collins, C. De Luca, A. Burrows, and L. Lipsitz. Age-related changes in open-loop and closed-loop postural control mechanisms. *Experimental brain research*, 104(3):480–492, 1995.

J. J. Collins and C. J. De Luca. Open-loop and closed-loop control of posture: a random-walk analysis of center-of-pressure trajectories. *Experimental brain research*, 95(2):308–318, 1993.

H. Corriveau, R. Hébert, M. Raiche, and F. Prince. Evaluation of postural stability in the elderly with stroke. *Archives of physical medicine and rehabilitation*, 85(7):1095–1101, 2004.

C. Cortes and M. Mohri. Auc optimization vs. error rate minimization. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

D. Cossock and T. Zhang. Subset ranking using regression. In H. Simon and G. Lugosi, editors, *Proceedings of COLT 2006*, volume 4005 of *Lecture Notes in Computer Science*, pages 605–619, 2006.

S. Couch, Z. Kazan, K. Shi, A. Bray, and A. Groce. Differentially private nonparametric hypothesis testing. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19, page 737–751. Association for Computing Machinery, 2019. ISBN 9781450367479.

G. Cumming. The new statistics: Why and how. *Psychological science*, 25(1):7–29, 2014.

X. H. Dang, B. Micenková, I. Assent, and R. T. Ng. Local outlier detection with interpretation. In *Machine Learning and Knowledge Discovery in Databases*, pages 304–320, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

D. Darling. The kolmogorov-smirnov, cramer-von mises tests. *The Annals of Mathematical Statistics*, 28(4):823–838, 1957.

J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

V. De la Pena and E. Giné. *Decoupling: from dependence to independence*. Springer Science and Business Media, 1999.

N. Deb and B. Sen. Multivariate rank-based distribution-free nonparametric testing using measure transportation, 2019.

N. Deb, B. B. Bhattacharya, and B. Sen. Efficiency lower bounds for distribution-free hotelling-type two-sample tests based on optimal transport, 2021.

R. Devore and G. Lorentz. *Constructive Approximation*. Springer, 1993.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

A. Di Bucchianico. Combinatorics, computer algebra and the wilcoxon-mann-whitney test. *Journal of Statistical Planning and Inference*, 79(2):349–364, 1999. ISSN 0378-3758.

M. D. Donsker. Justification and extension of doob's heuristic approach to the kolmogorov- smirnov theorems. *The Annals of Mathematical Statistics*, 23(2):277–281, 1952.

D. A. dos Santos, C. A. Fukuchi, R. K. Fukuchi, and M. Duarte. A data set with kinematic and ground reaction forces of human balance. *PeerJ*, 5:e3626, 2017.

R. Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, 1999.

A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator. *The Annals of Mathematical Statistics*, 27 (3):642 – 669, 1956.

M. Dwass. The Large-Sample Power of Rank Order Tests in the Two-Sample Problem. *The Annals of Mathematical Statistics*, 27(2):352 – 374, 1956.

J. Egan. *Signal Detection Theory and ROC Analysis*. Academic Press, 1975.

J. H. J. Einmahl and D. M. Mason. Generalized Quantile Processes. *The Annals of Statistics*, 20(2): 1062 – 1078, 1992.

c. Ertekin and C. Rudin. On equivalence relationships between classification and ranking algorithms. *Journal of Machine Learning Research*, 12:2905–2929, 2011.

M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.

D. Fanelli. How many scientists fabricate and falsify research? a systematic review and meta-analysis of survey data. *PLOS ONE*, 4(5):1–11, 2009.

T. Fawcett. An Introduction to ROC Analysis. *Pattern Recognition Letters*, 27:861–874, 2006.

R. J. Feise. Do multiple outcome measures require p-value adjustment? *BMC medical research methodology*, 2(1):8, 2002.

J. Feydy, T. Séjourné, F.-X. Vialard, S. ichi Amari, A. Trouvé, and G. Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences, 2018.

R. A. Fisher. Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 22(5):700–725, 1925.

J. Frery, A. Habrard, M. Sebban, O. Caelen, and L. He-Guelton. Efficient top rank optimization with gradient boosting for supervised anomaly detection. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD'17)*, Sept. 2017.

Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997. ISSN 0022-0000.

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

J. Friedman. On multivariate goodness-of-fit and two-sample testing. 03 2004.

J. H. Friedman and L. C. Rafsky. Multivariate Generalizations of the Wald-Wolfowitz and Smirnov Two-Sample Tests. *The Annals of Statistics*, 7(4):697 – 717, 1979.

J. C. Fu. Distribution theory of runs and patterns associated with a sequence of multi-state trials. *Statistica Sinica*, 6(4):957–974, 1996.

P. Gatev, S. Thomas, T. Kepple, and M. Hallett. Feedforward ankle strategy of balance during quiet stance in adults. *The Journal of physiology*, 514(3):915–928, 1999.

R. Genuer, J. Poggi, and C. Tuleau-Malot. Variable selection using random forests. *Pattern Recognition Letters*, 31(14):2225–2236, 2010.

J. D. Gibbons and S. Chakraborti. *Nonparametric Statistical Inference*, pages 977–979. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-04898-2. doi: 10.1007/978-3-642-04898-2_420.

E. Giné, R. Latała, and J. Zinn. Exponential and moment inequalities for u-statistics. In *High Dimensional Probability II*, pages 13–38. Birkhäuser Boston, 2000.

E. Giné and A. Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 38(6):907 – 921, 2002.

E. Giné and J. Zinn. Some limit theorems for empirical processes. *The Annals of Probability*, 12(4): 929–989, 2004.

E. Giné, V. Koltchinskii, and J. Zinn. Weighted uniform consistency of kernel density estimators. *The Annals of Probability*, 32(3B):2570–2605, 07 2004.

S. Girard and J. Saracco. An introduction to dimension reduction in nonparametric kernel regression. In *Regression methods for astrophysics*, volume 66 of *EAS Publications Series*, pages 167–196. EDP Sciences, 2014. doi: 10.1051/eas/1466012.

S. A. Glantz. Biostatistics: how to detect, correct and prevent errors in the medical literature. *Circulation*, 61(1):1–7, 1980.

N. Goix, A. Sabourin, and S. Clémençon. On anomaly ranking and excess-mass curves. In *AISTATS*, 2015.

O. Goldreich, S. Goldwasser, and D. Ron. Property testing and its connection to learning and approximation. 45(4):653–750, 1998.

J. Gottschall, J. Peinke, V. Lippens, and V. Nagel. Exploring the dynamics of balance data—movement variability in terms of drift and diffusion. *Physics Letters A*, 373(8-9):811–816, 2009.

J. Gou, A. Tamhane, D. Xi, and R. D. A class of improved hybrid hochberg-hommel type step-up multiple test procedures. *Biometrika*, 101(4):899–911, 2014.

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

A. Gretton, K. Fukumizu, Z. Harchaoui, and B. K. Sriperumbudur. A fast, consistent kernel two-sample test. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

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

A. Gretton, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu, and B. K. Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012b.

A. Gretton, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu, and B. K. Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems*, volume 25, 2012c.

B. Guedj and S. Robbiano. Pac-bayesian high dimensional bipartite ranking. *Journal of Statistical Planning and Inference*, 2018.

C. Gutenbrunner and J. Jurečková. Regression Rank Scores and Regression Quantiles. *The Annals of Statistics*, 20(1):305 – 330, 1992.

L. Györfi, M. Köhler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

J. Hajek. Some Extensions of the Wald-Wolfowitz-Noether Theorem. *The Annals of Mathematical Statistics*, 32(2):506 – 523, 1961.

J. Hájek. Asymptotically most powerful rank-order tests. *The Annals of Mathematical Statistics*, 33 (3):112–1147, 09 1962.

J. Hájek. Asymptotic normality of simple linear rank statistics under alternatives. *The Annals of Mathematical Statistics*, 39:325–346, 1968.

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

P. Hall and N. Tajvidi. Permutation tests for equality of distributions in high-dimensional settings. *Biometrika*, 89(2):359–374, 2002. ISSN 00063444.

M. Hallin and D. Paindaveine. Optimal tests for multivariate location based on interdirections and pseudo-Mahalanobis ranks. *The Annals of Statistics*, 30(4):1103 – 1133, 2002a.

M. Hallin and D. Paindaveine. Optimal procedures based on interdirections and pseudo-Mahalanobis ranks for testing multivariate elliptic white noise against ARMA dependence. *Bernoulli*, 8(6):787 – 815, 2002b.

M. Hallin and D. Paindaveine. Optimal rank-based tests for homogeneity of scatter. *The Annals of Statistics*, 36(3):1261 – 1298, 2008.

M. Hallin and M. L. Puri. Optimal Rank-Based Procedures for Time Series Analysis: Testing an ARMA Model Against Other ARMA Models. *The Annals of Statistics*, 16(1):402 – 432, 1988.

M. Hallin and M. L. Puri. Time series analysis via rank order theory: Signed-rank tests for arma models. *Journal of Multivariate Analysis*, 39(1):1–29, 1991. ISSN 0047-259X.

M. Hallin and O. Tribel. The efficiency of some nonparametric rank-based competitors to correlogram methods. *Lecture Notes-Monograph Series*, 35:249–262, 2000.

M. Hallin, D. L. Vecchia, and H. Liu. Center-outward r-estimation for semiparametric varma models, 2020.

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

P. R. Halmos. The Theory of Unbiased Estimation. *The Annals of Mathematical Statistics*, 17(1):34 – 43, 1946.

S. Hediger, L. Michel, and J. Naf. On the use of random forest for two-sample testing, 2021.

N. Henze. A multivariate two-sample test based on the number of nearest neighbor type coincidences. *The Annals of Statistics*, 16(2):772–783, 1988.

M. E. Hernandez, J. Snider, C. Stevenson, G. Cauwenberghs, and H. Poizner. A correlation-based framework for evaluating postural control stochastic dynamics. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 24(5):551–561, 2015.

C. Hill, B. Thompson, and E. Williams. A guide to conducting consensual qualitative research. *Counseling Psychologist*, 25:517–572, 10 1997.

A. Hinneburg and D. A. Keim. An efficient approach to clustering in large multimedia databases with noise. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, KDD'98, page 58–65. AAAI Press, 1998.

Y. Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4): 800–802, 1988.

J. L. Hodges. A Bivariate Sign Test. *The Annals of Mathematical Statistics*, 26(3):523 – 527, 1955.

J. L. Hodges and E. L. Lehmann. The Efficiency of Some Nonparametric Competitors of the t-Test. *The Annals of Mathematical Statistics*, 27(2):324 – 335, 1956.

P. Hodges, V. Gurfinkel, S. Brumagne, T. Smith, and P. Cordo. Coexistence of stability and mobility in postural control: evidence from postural compensation for respiration. *Experimental brain research*, 144(3):293–302, 2002.

W. Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19:293–325, 1948.

W. Hoeffding. The strong law of large numbers for u-statistics. *North Carolina State University. Dept. of Statistics*, 21:293–325, 1961.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. ISSN 01621459.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, pages 65–70, 1979.

G. Hommel. A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika*, 75(2):383–386, 1988.

F. B. Horak. Clinical measurement of postural control in adults. *Physical therapy*, 67(12):1881–1885, 1987.

F. B. Horak. Postural orientation and equilibrium: what do we need to know about neural control of balance to prevent falls? *Age and ageing*, 35(suppl_2):ii7–ii11, 2006.

C. Houdré and P. Reynaud-Bouret. Exponential inequalities, with constants, for u-statistics of order two. In *Stochastic Inequalities and Applications*, pages 55–69. Birkhäuser Basel, 2003.

H. Huang, K. G. Mehrotra, and C. K. Mohan. Outlier detection using modified-ranks and other variants. *Electrical Engineering and Computer Science - Technical Reports*, 72, 2011.

H. Huang, K. G. Mehrotra, and C. K. Mohan. Rank-based outlier detection. *Journal of Statistical Computation and Simulation*, 83:518 – 531, 2013.

Y. Ingster and I. Suslina. *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*, volume 169. Springer-Verlag Berlin Heidelberg, 2003.

Y. I. Ingster and I. A. Suslina. Minimax nonparametric hypothesis testing for ellipsoids and besov bodies. *ESAIM: Probability and Statistics*, 4:53–135, 2000.

J. P. A. Ioannidis. Why most published research findings are false. *PLOS Medicine*, 2(8), 08 2005.

S. Janitza and R. Hornung. On the overestimation of random forest's out-of-bag error. *PloS one*, 13 (8), 2018.

K. Järvelin and J. Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In N. Belkin, P. . Ingwersen, , and M.-K. Leong, editors, *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 41–48, 2000.

W. Jitkrittum, W. Xu, Z. Szabó, K. Fukumizu, and A. Gretton. A linear-time kernel goodness-of-fit test. In *Advances in Neural Information Processing Systems*, pages 262–271, 2017.

T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, page 133–142. Association for Computing Machinery, 2002.

T. Joachims. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, page 217–226. Association for Computing Machinery, 2006.

N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*, volume 1. John Wiley and Sons, 2nd edition, 1994.

M. Jones. The performance of kernel density functions in kernel distribution function estimation. *Statistics and Probability Letters*, 9(2):129–132, 1990.

J. Jurečková, J. Picek, and A. E. Saleh. Rank tests and regression rank score tests in measurement error models. *Computational Statistics and Data Analysis*, 54(12):3108–3120, 2010. ISSN 0167-9473.

K. Kafadar. EDITORIAL: Statistical significance, P-values, and replicability. *The Annals of Applied Statistics*, 15(3):1081 – 1083, 2021.

G. Kerr, C. Worringham, M. Cole, P. Lacherez, J. Wood, and P. Silburn. Predictors of future falls in parkinson disease. *Neurology*, 75(2):116–124, 2010.

A. Khasnis, R. Gokula, et al. Romberg's test. *Journal of postgraduate medicine*, 49(2):169, 2003.

H. Kiers, S. Brumagne, J. Van Dieen, P. van der Wees, and L. Vanhees. Ankle proprioception is not targeted by exercises on an unstable surface. *European journal of applied physiology*, 112(4): 1577–1585, 2012.

V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593 – 2656, 2006.

V. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. In *High Dimensional Probability II*, pages 443–457, Boston, MA, 2000.

V. Korolyuk and Y. Borovskich. *Theory of U-Statistics*, volume 273 of *Mathematics and Its Applications*. Springer Netherlands, 1994.

M. Korpelainen, J. Korpelainen, R. Sotaniemi, K. Virranniemi, and V. Myllylä. Postural sway and falls in parkinson's disease: a regression approach. *Movement Disorders*, 22(13):1927–1935, 2007.

G. Koshevoy and K. Mosler. Zonoid trimming for multivariate distributions. *The Annals of Statistics*, 25(5):1998–2017, 1997.

H. Koul. Some convergence theorems for ranks and weighted empirical cumulatives. *The Annals of Mathematical Statistics*, (41):1768–1773, 1970.

H. Koul. *Weighted Empirical Processes in Dynamic Nonlinear Models*, volume 166 of *Lecture Notes in Statistics*. Springer, 2nd edition, 2002.

W. Krzanowski and D. Hand. *ROC Curves for Continuous Data (1st ed.)*. 2009.

M. J. Kurz, D. J. Arpin, B. L. Davies, and R. Harbourne. The stochastic component of the postural sway variability is higher in children with balance impairments. *Annals of biomedical engineering*, 41(8):1703–1712, 2013.

D. Lafond. *Contribution à l'évaluation de l'équilibre quasi-statique à l'aide d'une plate-forme de force*. PhD thesis, 2003.

D. Lafond, M. Duarte, and F. Prince. Comparison of three methods to estimate the center of mass during balance assessment. *Journal of biomechanics*, 37(9):1421–1426, 2004.

J. Lam-Weil, B. Laurent, and J.-M. Loubes. Minimax optimal goodness-of-fit testing for densities under a local differential privacy constraint. *ArXiv*, abs/2002.04254, 2020.

J. Lam-Weil, A. Carpentier, and B. K. Sriperumbudur. Local minimax rates for closeness testing of discrete distributions. *Bernoulli*, 28(2):1179 – 1197, 2022.

H. Lamba and L. Akoglu. *Learning On-the-Job to Re-rank Anomalies from Top-1 Feedback*, pages 612–620. 2019.

T. A. Lang and D. G. Altman. Basic statistical reporting for articles published in biomedical journals: The "statistical analyses and methods in the published literature" or the sampl guidelines. *International Journal of Nursing Studies*, 52(1):5–9, 2015. ISSN 0020-7489.

M. Latt, S. Lord, J. Morris, and V. Fung. Clinical and physiological assessments for elucidating falls risk in parkinson's disease. *Movement Disorders*, 24:1280–1289, 2009.

M. Lauk, C. C. Chow, L. A. Lipsitz, S. L. Mitchell, and J. J. Collins. Assessing muscle stiffness from quiet stance in parkinson's disease. *Muscle and Nerve: Official Journal of the American Association of Electrodiagnostic Medicine*, 22(5):635–639, 1999.

J. Leach, M. Mancini, R. Peterka, T. Hayes, and F. Horak. Validating and calibrating the nintendo wii balance board to derive reliable center of pressure measures. *Sensors*, 14(10):18244–18267, 2014.

M. Ledoux. *The Concentration of Measure Phenomenon*. Mathematical surveys and monographs. American Mathematical Society, 2001.

A. J. Lee. *U-statistics: Theory and practice*. Marcel Dekker, Inc., New York, 1990.

E. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer, 2005.

E. L. Lehmann and F.-W. Scholz. Ancillarity. *Institute of Mathematical Statistics Lecture Notes - Monograph Series*, 17:32–51, 1992.

O. V. Lepski and V. G. Spokoiny. Minimax nonparametric hypothesis testing: The case of an inhomogeneous alternative. *Bernoulli*, 5(2):333–358, 1999.

C. Li, W. Jiang, and M. Tanner. General oracle inequalities for gibbs posterior with application to ranking. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 512–521. Proceedings of Machine Learning Research, 12–14 Jun 2013.

C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Poczos. Mmd gan: Towards deeper understanding of moment matching network. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Liu. On a notion of data depth based upon random simplices. *The Annals of Statistics*, 18(1):405–414, 1990.

F. Liu, K. Ting, and Z. Zhou. Isolation forest. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 413–422, 2008.

R. Y. Liu. Control charts for multivariate processes. *Journal of the American Statistical Association*, 90(432):1380–1387, 1995.

R. Y. Liu and K. Singh. A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association*, 88(421):252–260, 1993.

T.-Y. Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, mar 2009.

M. E. Lopes, L. Jacob, and M. Wainwright. A more powerful two-sample test in high dimensions using random projection. In *Advances in Neural Information Processing Systems*, 2011.

D. Lopez-Paz and M. Oquab. Revisiting classifier two-sample tests. 2016.

I. Loram. Postural control and sensorimotor integration. *Grieve's Modern Musculoskeletal Physiotherapy E-Book*, page 28, 2015.

I. Lovato, A. Pini, A. Stamm, and S. Vantini. Model-free two-sample test for network-valued data. *Computational Statistics & Data Analysis*, 144:106896, 2020. ISSN 0167-9473.

A. Lung-Yut-Fong, C. Lévy-Leduc, and O. Cappé. Homogeneity and change-point detection tests for multivariate data using rank statistics. *Journal de la société française de statistique*, 156(4): 133–162, 2015.

A. Mahboobin, P. J. Loughlin, and M. S. Redfern. A model-based approach to attention and sensory integration in postural control of older adults. *Neuroscience letters*, 429(2-3):147–151, 2007.

P. Major. An estimate on the supremum of a nice class of stochastic integrals and u-statistics. *Probability Theory and Related Fields*, 134(3):489–537, 2006.

M. Mancini, P. Carlson-Kuhta, C. Zampieri, J. Nutt, L. Chiari, and F. Horak. Postural sway as a marker of progression in parkinson's disease: a pilot longitudinal study. *Gait and posture*, 36(3): 471–476, 2012a.

M. Mancini, A. Salarian, P. Carlson-Kuhta, C. Zampieri, L. King, and L. Chiari. Isway: a sensitive, valid and reliable measure of postural control. *Journal of Neuroengineering and Rehabilitation*, 9, 2012b.

H. Mann and D. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18:50–60, 1947.

K. Masani, A. H. Vette, and M. R. Popovic. Controlling balance during quiet standing: proportional and derivative controller generates preceding motor command to body sway position observed in experiments. *Gait and posture*, 23(2):164–172, 2006.

P. Massart. The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality. *The Annals of Probability*, 18(3):1269 – 1283, 1990.

D. K. McClish. Analyzing a portion of the roc curve. *Medical Decision Making*, 9(3):190–195, 1989.

C. McDiarmid. *On the method of bounded differences*, page 148–188. London Mathematical Society Lecture Note Series. Cambridge University Press, 1989.

C. McDiarmid. *Concentration*, pages 195–248. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998. ISBN 978-3-662-12788-9. doi: 10.1007/978-3-662-12788-9_6.

I. Melzer, N. Benjuya, and J. Kaplanski. Postural stability in the elderly: a comparison between fallers and non-fallers. *Age and ageing*, 33(6):602–607, 2004.

A. Menon and R. Williamson. Bipartite ranking: A risk theoretic perspective. *Journal of Machine Learning Research*, 7:1–102, 2016.

S. L. Mitchell, J. Collin, C. J. De Luca, A. B. Burrows, and L. A. Lipsitz. Open-loop and closed-loop postural control mechanisms in parkinson's disease: increased mediolateral activity during quiet standing. *Neuroscience letters*, 197(2):133–136, 1995.

A. D. G. Moher D., Schulz K. F. The consort statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet*, 357(9263):1191–1194, 2001.

K. Mosler. Depth statistics. In C. Becker, R. Fried, and S. Kuhnt, editors, *Robustness and Complex Data Structures: Festschrift in Honour of Ursula Gather*, pages 17–34. Springer, 2013.

J. Möttönen, H. Oja, and J. Tienari. On the efficiency of multivariate spatial sign and rank tests. *The Annals of Statistics*, 25(2):542–552, 1997.

J. Möttönen, H. Oja, and R. Serfling. Multivariate generalized spatial signed-rank methods. *Journal of Statistical Research*, 39(1):19–42, 2005. ISSN 0256-422X.

J. Muir, D. Kiel, M. Hannan, J. Magaziner, and C. Rubin. Dynamic parameters of balance which correlate to elderly persons with a history of falls. *Plos One*, 8(8), 2013.

S. Mukherjee, D. Agarwal, N. R. Zhang, and B. B. Bhattacharya. Distribution-free multisample tests based on optimal matchings with applications to single cell genomics. *Journal of the American Statistical Association*, pages 1–12, 2020.

J. Möttönen and H. Oja. Multivariate spatial sign and rank methods. *Journal of Nonparametric Statistics*, 5(2):201–213, 1995.

E. Müller, I. Assent, P. Iglesias, Y. Mülle, and K. Böhm. Outlier ranking via subspace analysis in multiple views of the data. In *2012 IEEE 12th International Conference on Data Mining*, pages 529–538, 2012.

E. Müller, P. I. Sánchez, Y. Mülle, and K. Böhm. Ranking outlier nodes in subspaces of attributed graphs. In *2013 IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*, pages 216–222, 2013. doi: 10.1109/ICDEW.2013.6547453.

E. Nadaraya. Somenew estimates for distribution functions. *Theory of Probability and its Apllications*, 9(3):497–500, 1964.

H. Narasimhan and S. Agarwal. On the relationship between binary classification, bipartite ranking, and binary class probability estimation. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

H. Narasimhan and S. Agarwal. Support vector algorithms for optimizing the partial area under the roc curve. *Neural Computation*, 29:1919–1963, 2017.

N. Neumeyer. A central limit theorem for two-sample u-processes. *Statistics and Probability Letters*, 67(1):73 – 85, 2004.

K. Newell, S. Slobounov, E. Slobounova, and P. Molenaar. Stochastic processes in postural center-of-pressure profiles. *Experimental Brain Research*, 113(1):158–164, 1997.

R. T. Ng and J. Han. Efficient and effective clustering methods for spatial data mining. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, page 144–155. Morgan Kaufmann Publishers Inc., 1994.

A. Nicolaï and J. Audiffren. Model-space regularization and fully interpretable algorithms for postural control quantification. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, volume 2, pages 177–182. IEEE, 2018.

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

H. Oja. Descriptive statistics for multivariate distributions. *Statistics and Probability Letters*, 1(6): 327–332, 1983.

H. Oja. *Multivariate Nonparametric Methods with R: An approach based on spatial signs and ranks*. Springer-Verlag New York, 2010.

G. Pagnacco, E. Oggero, and C. Wright. Biomedical instruments versus toys: a preliminary comparison of force platforms and the nintendo wii balance board. *Biomedical Sciences Instrumentation*, 47:12–17, 2011.

R. Palmieri, C. Ingersoll, M. Stone, and B. Krause. Center-of-pressure parameters used in the assessment of postural control. *Journal of Sport Rehabilitation*, 11(1):51–66, 2002.

S. Park, J. M. Goo, and C. Jo. Receiver operating characteristic (roc) curve: Practical review for radiologists. 2004.

M. Patel, P.-A. Fransson, D. Lush, and S. Gomez. The effect of foam surface properties on postural stability assessment while standing. *Gait and posture*, 28(4):649–656, 2008.

L. Perini, C. Galvin, and V. Vercruyssen. *A Ranking Stability Measure for Quantifying the Robustness of Anomaly Detection Methods*. Springer International Publishing, 2020.

T. Perneger. What's wrong with bonferroni adjustments. *British Medical Journal*, 316(7139):1236–1238, 1998a.

T. V. Perneger. What's wrong with bonferroni adjustments. *British Medical Journal*, 316(7139): 1236–1238, 1998b. ISSN 0959-8138.

P. P. Perrin, C. Jeandel, C. A. Perrin, and M. C. Bene. Influence of visual control, conduction, and central integration on static and dynamic balance in healthy older adults. *Gerontology*, 43(4): 223–231, 1997.

R. J. Peterka. Sensorimotor integration in human postural control. *Journal of neurophysiology*, 88 (3):1097–1118, 2002.

W. Polonik. Minimum volume sets and generalized quantile processes. *Stochastic Processes and their Applications*, 69(1):1–24, 1997. ISSN 0304-4149.

T. E. Prieto, J. B. Myklebust, R. G. Hoffmann, E. G. Lovett, and B. M. Myklebust. Measures of postural steadiness: differences between healthy young and elderly adults. *IEEE Transactions on biomedical engineering*, 43(9):956–966, 1996.

J. T. Præstgaard. Permutation and bootstrap kolmogorov-smirnov tests for the equality of two distributions. *Scandinavian Journal of Statistics*, 22(3):305–322, 1995.

M. Puri and P. Sen. *Nonparametric Methods in Multivariate Analysis*. Nonparametric Methods in Multivariate Analysis. Wiley, 1993.

J. Qian, J. Root, V. Saligrama, and Y. Chen. A rank-svm approach to anomaly detection, 2014.

F. Quijoux, A. Nicolaï, I. Chairi, I. Bargiotas, D. Ricard, A. Yelnik, L. Oudre, F. Bertin-Hugault, P. P. Vidal, N. Vayatis, S. Buffat, and J. Audiffren. A review of center of pressure (cop) variables to quantify standing balance in elderly people: Algorithms and open-access code. *Physiological reports*, 9(22), 11 2021.

A. Rachev. *Probability Metrics and the Stability of Stochastic Models*. Wiley, 1991.

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

A. Ramdas, N. Garcia, and M. Cuturi. On wasserstein two sample testing and related families of nonparametric tests, 2015.

A.-S. Richard. Model comparisons and r2. *The American Statistician*, 48(2):113–117, 1994.

S. Rinalduzzi, C. Trompetto, L. Marinelli, A. Alibardi, P. Missori, and F. Fattapposta. Balance dysfunction in parkinson's disease. *BioMed Research International*, 2015.

T. Ronkainen, H. Oja, and P. Orponen. Computation of the multivariate oja median. In *Developments in Robust Statistics*, pages 344–359. Physica-Verlag HD, 2003.

L. Z. Rubenstein. Falls in older people: epidemiology, risk factors and strategies for prevention. *Age and ageing*, 35(suppl_2):ii37–ii41, 2006.

R. Rubinfeld and M. Sudan. Robust characterizations of polynomials with applications to program testing. *SIAM Journal on Computing*, 25(2):252–271, 1996.

C. Rudin. Ranking with a P-Norm Push. In H. Simon and G. Lugosi, editors, *Proceedings of COLT 2006*, volume 4005 of *Lecture Notes in Computer Science*, pages 589–604, 2006.

C. Rudin and R. E. Schapire. Margin-based ranking and an equivalence between adaboost and rankboost. *Journal of Machine Learning Research*, 10(77):2193–2232, 2009.

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

A. Sandholm, N. Pronost, and D. Thalmann. Motionlab: A matlab toolbox for extracting and processing experimental motion capture data for neuromuscular simulations. In *3D Physiological Human Workshop*, pages 110–124. Springer, 2009.

M. F. Schilling. Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association*, 81(395):799–806, 1986.

B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, 2002.

B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt. Support vector method for novelty detection. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, page 582–588. MIT Press, 1999.

B. Schölkopf, J. Platt, A. J. Shawe-Taylor, J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7), 2001.

B. Schölkopf, K. Tsuda, and J.-P. Vert. *Kernel Methods in Computational Biology*. MIT Press, 2003.

A. Schrab, I. Kim, M. Albert, B. Laurent, B. Guedj, and A. Gretton. Mmd aggregated two-sample test, 2021.

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

R. Serfling. *Approximation theorems of mathematical statistics*. John Wiley and Sons, 1980.

D. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures (5th ed.)*. Chapman and Hall/CRC, 2011.

G. Shorack and J. Wellner. *Empirical Processes with Applications to Statistics*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 2009.

N. Si, K. Murthy, J. Blanchet, and V. A. Nguyen. Testing group fairness via optimal transport projections. *ArXiv*, abs/2106.01070, 2021.

N. Smirnov. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bulletin Mathématique de l'Université de Moscou*, 2, 1939.

C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

R. Srivastava, P. Li, and D. Ruppert. Raptt: An exact two-sample test in high dimensions using random projections. *Journal of Computational and Graphical Statistics*, 25(3):954–970, 2016.

I. Steinwart, D. Hush, and C. Scovel. A classification framework for anomaly detection. *Journal of Machine Learning Research*, 6(8):211–232, 2005.

D. A. Sterling, J. A. O'connor, and J. Bonadies. Geriatric falls: injury severity is high and disproportionate to mechanism. *Journal of Trauma and Acute Care Surgery*, 50(1):116–119, 2001.

J. Stevens, P. Corso, E. Finkelstein, and M. TR. The costs of fatal and non-fatal falls among older adults. *Injury prevention : journal of the International Society for Child and Adolescent Injury Prevention*, 12(5):290–295, 2006.

W. Stute. Conditional U-Statistics. *The Annals of Probability*, 19(2):812 – 825, 1991.

D. Sutherland, H. Tung, H. Strathmann, S. De, A. Ramdas, and S. A. Generative models and model criticism via optimized maximum mean discrepancy. *International Conference on Learning Representations*, 2017.

J. Swanenburg, E. de Bruin, D. Uebelhart, and T. Mulder. Falls prediction in elderly people: a 1-year prospective study. *Gait and Posture*, 31(3):317–321, 2010.

J. A. Swets. Measuring the accuracy of diagnostic systems. *Science*, 240(4857):1285–1293, 1988.

G. Szekely. E-statistics: The energy of statistical samples. 01 2003.

G. Szekely and M. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5, 11 2004.

G. J. Székely and M. L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013. ISSN 0378-3758.

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

M. Talagrand. Sharper Bounds for Gaussian and Empirical Processes. *The Annals of Probability*, 22 (1):28 – 76, 1994.

M. Talagrand. A new look at independence. *The Annals of Probability*, 24(1):1 – 34, 1996a.

T. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. 81:73–205, 1995. ISSN 16181913.

T. Talagrand. New concentration inequalities in product spaces. *Inventiones mathematicae*, 126: 505–563, 1996b. ISSN 14321297.

Y. Tawaki and T. Murakami. Evaluation of langevin model for human stabilogram based on reproducibility of statistical indicators. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1934–1939. IEEE, 2019.

M. Thiese, Z. Arnold, and S. Walker. The misuse and abuse of statistics in biomedical research. *Biochemia Medica*, 25(1):5–11, 2015.

A. Thomas, S. Clemencon, A. Gramfort, and A. Sabourin. Anomaly Detection in Extreme Regions via Empirical MV-sets on the Sphere. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1011–1019. PMLR, 2017.

M. E. Tinetti. Preventing falls in elderly persons. *New England Journal of Medicine*, 348(1):42–49, 2003a.

M. E. Tinetti. Preventing falls in elderly persons. *New England journal of medicine*, 348(1):42–49, 2003b.

J. W. Tukey. Mathematics and the picturing of data. In R. D. James, editor, *Proceedings of the International Congress of Mathematicians*, volume 2, pages 523–531. Canadian Mathematical Congress, 1975.

Y. Um and R. H. Randles. Nonparametric tests for the multivariate multi-sample location problem. *Statistica Sinica*, 8(3):801–812, 1998.

L. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984.

S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

A. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

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

M. Van Diest, J. Lamoth, C, J. Stegenga, G. J. Verkerke, and K. Postema. Exergaming for balance training of elderly: state of the art and future developments. *Journal of neuroengineering and rehabilitation*, 10(1):101, 2013.

V. Vapnik. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann, 1992.

V. N. Vapnik and A. Y. Chervonenkis. *On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities*. Springer International Publishing, Cham, 2015. ISBN 978-3-319-21852-6. doi: 10.1007/978-3-319-21852-6_3.

Y. Vardi and C.-H. Zhang. The multivariate l_1-median and associated data depth. *Proceedings of the National Academy of Sciences*, 97(4):1423–1426, 2000.

C. Villani. *Optimal Transport*. Springer-Verlag Berlin Heidelberg, 2009.

E. H. W. Kotlowski, K. Dembczynski. Bipartite ranking through minimization of univariate loss. In *International Conference on Machine Learning*, pages 1113–1120, 2011.

A. Wald and J. Wolfowitz. On a Test Whether Two Samples are from the Same Population. *The Annals of Mathematical Statistics*, 11(2):147 – 162, 1940.

S. Walter. The partial area under the summary roc curve. *Statistics in medicine*, 24:2025–40, 07 2005.

M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019*, pages 692–702. IEEE, 2019.

Y. Wang, K. Li, and S. Gan. A kernel connectivity-based outlier factor algorithm for rare data detection in a baking process. *10th IFAC Symposium on Advanced Control of Chemical Processes ADCHEM 2018*, 51(18):297–302, 2018. ISSN 2405-8963.

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

D. A. Winter, A. E. Patla, F. Prince, M. Ishac, and K. Gielo-Perczak. Stiffness control of balance in quiet standing. *Journal of neurophysiology*, 80(3):1211–1221, 1998.

D. A. Winter, A. E. Patla, S. Rietdyk, and M. G. Ishac. Ankle muscle stiffness in the control of balance during quiet standing. *Journal of neurophysiology*, 85(6):2630–2633, 2001.

J. Wolfowitz. On the Theory of Runs with some Applications to Quality Control. *The Annals of Mathematical Statistics*, 14(3):280 – 288, 1943.

J. Wood, N. Freemantle, M. King, and N. I. Trap of trends to statistical significance: likelihood of near significant p value becoming more significant with extra data. *British Medical Journal*, 348, 2014.

H. Yang, K. Lu, X. Lyu, and F. Hu. Two-way partial auc and its properties. *Statistical Methods in Medical Research*, 28(1):184–195, 2019.

V. M. Zatsiorsky and M. Duarte. Instant equilibrium point and its migration in standing tasks: rambling and trembling components of the stabilogram. *Motor control*, 3(1):28–38, 1999.

V. M. Zatsiorsky and V. M. Zaciorskij. *Kinetics of human motion*. Human Kinetics, 2002.

A. Zecevic, A. Salmoni, M. Speechley, and A. Vandervoort.

W.-X. Zhou, C. Zheng, and Z. Zhang. Two-sample smooth tests for the equality of distributions. *Bernoulli*, 23(2):951 – 989, 2017.

Y. Zuo and R. Serfling. General notions of statistical depth function. *The Annals of Statistics*, 28(2): 461–482, 2000.

A | Generalized two-sample R-processes and efficient two-sample tests

Abstract. This short chapter analyzes generalized R-processes, insofar alternative assumptions are considered for the score-generating functions from Chapter 5. Although those results are fundamental, they do not apply for discontinuous choices of ϕ such as the ones introduced in the works of Boyd et al. (2012); Clémençon and Vayatis (2007); Cossock and Zhang (2006). Also, we present an adaptive two-sample homogeneity test based on R-statistics aiming to optimize the procedure proposed in Chap. 6. It relies on aggregating a sequence of linear R-statistics when indexed by a class of score-generating functions.

A.1 General score-generating functions

Chapter 5 introduced a class of two-sample linear rank statistics as W-ranking performance measure related to a given score-generating function $\phi(u)$, a class of scoring functions $\mathscr{S}$ and based on the independent *i.i.d.* samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$. In particular, we studied concentration inequalities and the performance of empirical maximizers when $\phi(u)$ is assumed to be nondecreasing and twice continuously differentiable. We aim here at proving similar results under mild assumptions on the score-generating function and deriving concentration result of the class of score-generating functions, based on the work of Koul (2002) (Chapter 3) and Clémençon and Vayatis (2007).

We recall that the *r.v.* $\mathbf{X}$ and $\mathbf{Y}$ are independent and defined on the same probability space $(\Omega, \mathscr{F}, \mathbb{P})$, valued in the same space $\mathscr{X}$, *resp.* drawn from G and H. Let $p \in (0,1)$ be the 'theoretical' fraction of positive instances. For $N \geq 1/p$, we suppose that $n = \lfloor pN \rfloor$ and $m = \lceil (1-p)N \rceil = N - n$. Define the mixture probability distribution $F = pG + (1-p)H$. For any $s \in \mathscr{S}$, the distribution of $s(\mathbf{X})$ (*i.e.* the image of G by s) is denoted by G_s, that of $s(\mathbf{Y})$ (*i.e.* the image of H by s) by H_s. We also denote by F_s the image of distribution F by s. For simplicity, the same notations are used to mean the related cumulative distribution functions. We also introduce their statistical versions $\widehat{G}_{s,n}(t) = (1/n)\sum_{i=1}^{n} \mathbb{I}\{s(\mathbf{X}_i) \leq t\}$ and $\widehat{H}_{s,m}(t) = (1/m)\sum_{j=1}^{m} \mathbb{I}\{s(\mathbf{Y}_j) \leq t\}$ and define

$$\widehat{F}_{s,N}(t) = (n/N)\widehat{G}_{s,n}(t) + (m/N)\widehat{H}_{s,m}(t)\ . \tag{A.1.1}$$

Since $n/N \to p$ as N tends to infinity, the quantity above is a natural estimator of the *c.d.f.* F_s. We suppose the following assumptions fulfilled.

***Assumption* 10.** *$\phi \in \mathscr{C}$ where $\mathscr{C} := \{\phi : [0,1] \to \mathbb{R},\ \phi$ cadlag, $\|\phi\|_{tv} := \phi(1) - \phi(0) = 1\}$ and the set of non-continuity points is countable.*

In particular, denoting $\xi_1, \ldots, \xi_J$, $J \in \mathbb{N}^*$ the sequence of discontinuity points of the function ϕ, such that, for all $j \leq J-1$, ϕ is continuous increasing function on the open interval (ξ_j, ξ_{j+1}). Therefore, for all functions ℓ defined on a given ξ_j, one has

$$\int_{[\xi_j, \xi_j]} \ell d\phi = \ell(\xi_j)[\phi(\xi_j^+) - \phi(\xi_j^-)].$$

Let a subclass $\mathscr{S}_0$ of $\mathscr{S}$ such that the following assumptions are fulfilled.

***Assumption* 11.** *The class of scoring functions $\mathscr{S}_0 \subset \mathscr{S}$ defines a* VC*-class of finite* VC*-dimension $\mathscr{V}$.*

***Assumption* 12.** *The optimal element s^* exists and lies in $\mathscr{S}$. The empirical optimal function of $\mathscr{S}_0$ is denoted by $\hat{s}_N$.*

***Assumption* 13.** *Let $M > 0$. For all $s \in \mathscr{S}_0$, the random variables $s(\mathbf{X})$, $s(\mathbf{Y})$ are continuous, with density functions that are twice differentiable and have Sobolev $\mathscr{W}^{2,\infty}$-norms[1] bounded by $M < +\infty$.*

***Assumption* 14.** *Let $m > 0$. For all $s \in \mathscr{S}_0$, the probability density functions of the random variables $s(\mathbf{X})$, $s(\mathbf{Y})$ are strictly bounded above $g_s(t) > m$, $h_s(t) > m$, for all $t \in \mathbb{R}$.*

[1] Recall that the Sobolev space $\mathscr{W}^{2,\infty}$ is the space of all Borelian functions $h : \mathbb{R} \to \mathbb{R}$ such that h and its first and second order weak derivatives h' and h'' are bounded almost-everywhere. Denoting by $||.||_\infty$ the norm of the Lebesgue space L_∞ of Borelian and essentially bounded functions, $\mathscr{W}^{2,\infty}$ is a Banach space when equipped with the norm $||h||_{2,\infty} = \max\{||h||_\infty, ||h'||_\infty, ||h''||_\infty\}$.

Definition 82. *The two-sample 'W_ϕ-ranking performance measure' is defined, for all score-generating function $\phi \in \mathscr{C}$ and for all scoring function $s \in \mathscr{S}$, as the functional:*

$$W(\phi, s) = \mathbb{E}[(\phi \circ F_s)(s(\mathbf{X}))] \, . \tag{A.1.2}$$

Its empirical counterpart is expressed as a R-process, as follows:

$$\widehat{W}_{n,m}(\phi, s) = \sum_{i=1}^{n} \phi\left(\frac{\text{Rank}(s(\mathbf{X}_i))}{N+1}\right) \, , \tag{A.1.3}$$

where $\text{Rank}(t) := N\widehat{F}_{s,N}(t) = \sum_{i \le n} \mathbb{I}\{s(\mathbf{X}_i) \le t\} + \sum_{j \le m} \mathbb{I}\{s(\mathbf{Y}_j) \le t\}$ is the rank statistic and defined for all $s \in \mathscr{S}$. Optimality in the sense of (cite) is with respect to the element $s \in \mathscr{S}_0$ that maximizes the empirical W-ranking performance measure for a given score-generating function ϕ, *i.e.* ,

$$\widehat{s}_{n,m} \in \arg\max_{s \in \mathscr{S}_0} \widehat{W}_{n,m}(\phi, s) \, . \tag{A.1.4}$$

Considering a subclass $\mathscr{C}_0$ of $\mathscr{C}$, we aim at providing similar nonasymptotic bounds *w.r.t.* to the optimal element $s(z)$ and as well as the optimal $\phi(u)$. First, introduce the two-sample stochastic process $\widehat{Z}_n(t, s)$, for all $(t, s) \in [0,1] \times \mathscr{S}_0$, through

$$\widehat{Z}_n(t, s) = \sum_{i=1}^{n} \mathbb{I}\{\text{Rank}(s(\mathbf{X}_i)) \le Nt\} \, . \tag{A.1.5}$$

Let $\phi \in \mathscr{C}$ satisfy the Assumption (10), by performing integration by parts and linear change of variables yields

$$\widehat{W}_N(\phi, s) = \int \phi\left(\frac{Nt}{N+1}\right) \widehat{Z}_n(dt, s) = \phi(1)n - \int_0^1 \widehat{Z}_n\left(\frac{(N+1)x}{N}, s\right) d\phi(x) \, . \tag{A.1.6}$$

Hence, we will study the process $\widehat{Z}_n(t, s)$ and in particular its fluctuations through uniform deviation bounds on $[0,1] \times \mathscr{S}_0$ in order to obtain uniform bounds on $\mathscr{C}_0 \times \mathscr{S}_0$ for the process $\widehat{W}_N(\phi, s)$.

Remark 11. *Notice that for $\phi(u) = u$, the equality is trivially obtained.*

Uniform and Linear Approximation of the Two-Sample Process $\widehat{Z}_n$. In order to study the fluctuations of the process to obtain statistical guarantees, we need to linearize its structure due to the non-independent sum. Classic tools such as orthogonal projections onto linear combinations of *i.i.d.* variables cannot be directly applied and we follow the footsteps of Koul (2002) by considering the equivalent process $\widehat{S}_n(t, s)$ to exhibit the leading empirical process of order $O_{\mathbb{P}}(N^{-1/2})$, defined as follows

$$\widehat{S}_n(t, s) = \frac{1}{n} \sum_{i \le n} \mathbb{I}\{s(\mathbf{X}_i) \le \widehat{F}^{-1}_{N,s}(t)\} \, , \tag{A.1.7}$$

for all (t, s). It is an estimator of the following empirical process S_n of mean S:

$$S_n(t, s) = \frac{1}{n} \sum_{i \le n} \mathbb{I}\{s(\mathbf{X}_i) \le F_s^{-1}(t)\} \tag{A.1.8}$$

$$S(t, s) = \mathbb{E}[\mathbb{I}\{s(\mathbf{X}) \le F_s^{-1}(t)\}] = G_s(F_s^{-1}(t)) \tag{A.1.9}$$

The following result establishes a linear approximation of the stochastic process $\widehat{S}_n(t, s)$ through the sum of the function $S(t, s)$ and a leading term $\widehat{T}_{n,m}(t, s)$ averaging *i.i.d.* variables that is of order

$O_{\mathbb{P}}(N^{-1/2})$ plus a remainder of higher order. As introduced by Koul (2002), we link the process $\widehat{S}_n(t,s)$ to the empirical process $S_n(t,s)$ by evaluating the latter at $F_s \circ \widehat{F}_{N,s}^{-1}(t)$. It consequently bridges the gap of $\widehat{S}_n(t,s)$'s correlated structure, at the price of an approximation error in the linearization.

Proposition 83. *Let $\delta \in (0,1)$. Consider the Assumptions 11, 13, 14. Therefore, for all scoring function $s \in \mathscr{S}_0$ and $t \in [0,1]$, the following decomposition holds*

$$\widehat{S}_n(t,s) = S(t,s) + \widehat{T}_{n,m}(t,s) + \mathscr{R}_{n,m}(t,s) \ , \tag{A.1.10}$$

with

$$\widehat{T}_{n,m}(t,s) = t\partial_t S(t,s) - S(t,s) + \frac{1}{n}\sum_{i=1}^{n}(1 - \frac{n}{N}\partial_t S(t,s))\mathbb{I}\{s(\mathbf{X}_i) \leq F_s^{-1}(t)\} \\ - \frac{1}{N}\partial_t S(t,s)\sum_{j=1}^{m}\mathbb{I}\{s(\mathbf{Y}_j) \leq F_s^{-1}(t)\} \ , \tag{A.1.11}$$

where the derivative of $S(t,s)$ w.r.t. $t \in [0,1]$ is equal to

$$\partial_t S(t,s) = \left(\frac{g_s}{f_s}\right)(F_s^{-1}(t))$$

and with probability $1-\delta$, the remainder process $\mathscr{R}_{n,m}(t,s)$ is of order $O_{\mathbb{P}}(N^{-1})$.

PROOF. Suppose the Assumptions of Proposition 83 fulfilled. Let $s \in \mathscr{S}_0$ and $t \in [0,1]$, we have

$$\widehat{S}_n(t,s) - S(t,s) = S_n(F_s \circ \widehat{F}_{N,s}^{-1}(t),s) - S(t,s) \\ = \widehat{V}_n(F_s \circ \widehat{F}_{N,s}^{-1}(t),s) + S(F_s \circ \widehat{F}_{N,s}^{-1}(t),s) - S(t,s) \ . \tag{A.1.12}$$

with the centered empirical process $\widehat{V}_n(t,s) = S_n(t,s) - S(t,s)$. The proof of the proposition relies on the following two lemmas.

Lemma 84. *Let $\delta \in (0,1)$, $(t,s) \in [0,1] \times \mathscr{S}_0$. Consider the Assumptions 11 and 13 satisfied, therefore there exists a nonnegative constant C, such that*

$$\mathbb{P}\left\{|\widehat{V}_n(F_s \circ \widehat{F}_{N,s}^{-1}(t),s) - \widehat{V}_n(t,s)| < x\right\} \geq 1 - \delta \ , \tag{A.1.13}$$

with $x = C\log(1/\delta)/N$.

Lemma 85. *Let $(t,s) \in [0,1] \times \mathscr{S}_0$. Consider the Assumptions 11, 13, 14 satisfied, therefore*

$$S(F_s \circ \widehat{F}_{N,s}^{-1}(t),s) = S(t,s) - (F_s \circ \widehat{F}_{N,s}^{-1}(t) - t)\partial_t S(t,s) + O_{\mathbb{P}}(N^{-1}) \ . \tag{A.1.14}$$

□

PROOF.Proof of Lemma 84. Let $\delta \in (0,1)$, $x, \varepsilon > 0$ and $t \in [0,1]$, . Consider the Assumptions 11 and 13 satisfied. Following the footsteps of Clémençon and Vayatis (2007), conditionally on the event $A(s,\varepsilon) = \{|F_s \circ \widehat{F}_{N,s}^{-1}(t) - t| < \varepsilon\}$, we have

$$|\widehat{V}_n(F_s \circ \widehat{F}_{N,s}^{-1}(t),s) - \widehat{V}_n(t,s)| \leq \sup_{t':|t-t'|<\varepsilon} |\widehat{V}_n(t',s) - \widehat{V}_n(t,s)|$$

Consider the distance $d_{i,s}$, for all $i \leq n$, defined by

$$d_{i,s} : (t,t') \mapsto \mathbb{I}\{s(\mathbf{X}_i) \in [F_s^{-1}(\max(t,t')), F_s^{-1}(\min(t,t'))]\} + |t-t'|$$

such that

$$|\widehat{V}_n(t',s) - \widehat{V}_n(t,s)| \leq \frac{1}{n}\sum_{i=1}^{n} d_{i,s}(t,t') =: \hat{d}_{n,s}(t,t')$$

where $\hat{d}_s$ is the weighted distance over $\mathbb{R}$ for a given $s(z)$. Then the the uniform bound of the distance on the event $A(s,\varepsilon)$ *w.r.t.* t' is equal to

$$\begin{aligned} \hat{D}(\varepsilon) &:= \sup_{t':|t-t'|<\varepsilon} \hat{d}_s(t,t') \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{s(\mathbf{X}_i) \in [F_s^{-1}(t-\varepsilon)), F_s^{-1}(t+\varepsilon))]\} + \varepsilon \end{aligned}$$

Therefore following the footsteps of Clémençon and Vayatis (2007) and using the Lemma 8.5 from van de Geer (2000), the result is obtained.

□

PROOF.Proof of Lemma 85. Let $(t,s) \in [0,1] \times \mathscr{S}_0$. Consider the Assumptions 11, 13, 14 satisfied. Note that

$$|F_s \circ \widehat{F}_{N,s}^{-1}(t) - t| \leq \sup_t |F_s(t) - \widehat{F}_{N,s}(t)| + 1/N \,,$$

implies $|F_s \circ \widehat{F}_{N,s}^{-1}(t) - t|$ is of order $O_{\mathbb{P}}(N^{-1})$, and by pointwise differentiability of the *c.d.f.* s G_s and H_s, a first order approximation of $S(F_s \circ \widehat{F}_{N,s}^{-1}(t), s)$ at t concludes on the result.

□

A.2 Adaptive two-sample homogeneity rank tests

This section presents an adaptive two-sample homogeneity rank test in the continuity of Chapter 6. By considering a sequence of R-statistics indexed by a class of score-generating functions, we propose a procedure intending to minimize both statistical errors by aggregating the former sequence. It extends the procedure proposed 6.2 wherein the bipartite ranking step (*Step 1*) is replaced by the exact maximization of the W_ϕ-performance criterion, thus depending on the choice of ϕ. We first detail the new version of the procedure and then outline the aggregated test statistic.

Two-stage homogeneity rank-based testing. The considered framework for this method is similar to the one of Chap. 6. Based on the observation of two independent *i.i.d.* samples $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ with $n, m \geq 1$, the goal is to test at level $\alpha \in (0,1)$, the homogeneity hypothesis:

$$\mathscr{H}_0 : \; W_\phi^* = \int_0^1 \phi(u)du \; \textit{versus} \; \mathscr{H}_1 : \; W_\phi^* > \int_0^1 \phi(u)du \,, \tag{A.2.1}$$

where the W_ϕ-criterion is defined in Def. A.1.2 and the considered class of two-sample linear rank statistic is defined by

$$\widehat{W}_{n,m}(s,\phi) = \frac{1}{n}\sum_{i=1}^{n} \phi\left(\frac{\text{Rank}(s(\mathbf{X}_i))}{N+1}\right) , \tag{A.2.2}$$

with $N = n+m$, see Eq. (A.1.3). Therefore, the two-stage procedure can be summarized as in the Figure A.1.

GENERALIZED TWO-SAMPLE RANK TESTS

Input. Two independent and *i.i.d.* samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$ of sizes $n, m \geq 2$ and valued in $\mathscr{X}$ - subsample sizes $n' < n$ and $m' < m$ - $\mathscr{A}$ algorithm operating on the class $\mathscr{S}_0$ of scoring functions on $\mathscr{X}$ - ϕ score-generating function - univariate two-sample rank test Φ_α^ϕ of level $\alpha \in (0,1)$

Two-split trick. Divide each of the original samples into two subsamples:

$$\{\mathbf{X}_1, \ldots, \mathbf{X}_{n'}\} \cup \{\mathbf{X}_{1+n'}, \ldots, \mathbf{X}_n\} \text{ and } \{\mathbf{Y}_1, \ldots, \mathbf{Y}_{m'}\} \cup \{\mathbf{Y}_{1+m'}, \ldots, \mathbf{Y}_m\}$$

1. **Maximizing the W_ϕ-ranking performance criterion.** Run algorithm $\mathscr{A}$ based on training data $\mathscr{D}_{n',m'} = \{\{\mathbf{X}_1, \ldots, \mathbf{X}_{n'}\}, \{\mathbf{Y}_1, \ldots, \mathbf{Y}_{m'}\}\}$, producing the scoring function

$$\widehat{s}(z) = \underset{s \in \mathscr{S}_0}{\arg\max}\, \widehat{W}_{n',m'}(\phi, s) \tag{A.2.3}$$

2. **Univariate rank test.** Form the univariate samples

$$\{\widehat{s}(\mathbf{X}_{1+n'}), \ldots, \widehat{s}(\mathbf{X}_n)\} \text{ and } \{\widehat{s}(\mathbf{Y}_{1+m'}), \ldots, \widehat{s}(\mathbf{Y}_m)\},$$

the outcome of the test being finally determined by computing the quantity

$$\Phi_\alpha^\phi\left(\{\widehat{s}(\mathbf{X}_{1+n'}), \ldots, \widehat{s}(\mathbf{X}_n)\} \cup \{\widehat{s}(\mathbf{Y}_{1+m'}), \ldots, \widehat{s}(\mathbf{Y}_m)\}\right) , \tag{A.2.4}$$

depending on $\{\text{Rank}(\widehat{s}(\mathbf{X}_{1+n'}), \ldots, \text{Rank}(\widehat{s}(\mathbf{X}_n))\}$, where $\text{Rank}(t) = \sum_{i=1+n'}^{n} \mathbb{I}\{\widehat{s}(\mathbf{X}_i) \leq t\} + \sum_{i=1+m'}^{m} \mathbb{I}\{\widehat{s}(\mathbf{Y}_j) \leq t)\}$.

Figure A.1. Extension of the ranking-based two-sample rank test procedure.

Remark 12. (ON THE CONSISTENCY OF THE PROCEDURE) *We can consider two possible assumptions for the score-generating functions to obtain the consistency of Proc. A.1. Under the framework of Chap. 6, if ϕ satisfies Assumption 9(i) (or 4 in Chap. 5), then one obtains nonasymptotic consistency of the method, for which all generalization results inherited from Chap. 5 are applicable. Recall that it supposes $\phi : [0,1] \mapsto \mathbb{R}$ to be nondecreasing and twice continuously differentiable. Under mild conditions on ϕ of Assumption 10, Section A.1, asymptotic consistency can be obtained thanks to the linearization result of Prop. 83.*

Remark 13. (ON THE ALGORITHM $\mathscr{A}$) *The procedure of Chap. 6 relies on bipartite ranking algorithms $\mathscr{A}$, where one minimizes the related empirical loss, see Chap 2, section 2.2. However,*

the present procedure aims to maximize the exact empirical version of the W_ϕ-ranking performance criterion. It is, therefore, possible to apply the deterministic gradient ascent algorithm introduced in Chap. 7 for semiparametric classes of scoring functions $\mathscr{S}_0$. Importantly, this allows for very interpretable high-dimensional testing procedures. We refer to some preliminary numerical results presented in the Conclusion part of the manuscript.

Aggregating multiple *R*-statistics. We now discuss the approach of Proc. A.1 wherein *Step 1* exactly maximizes the empirical W_ϕ-performance criterion based on $\mathscr{D}_{n',m'}$, and as expressed in Eq. (A.1.3). The goal is to construct a procedure that maximizes the power of the two-sample test by aggregating *R*-statistics for various choices of ϕ.

Consider $\mathscr{C}$ to be composed of $B \in \mathbb{N}^*$, finite, score-generating functions $\phi^{(1)}, \ldots, \phi^{(B)}$. Refer to Remark 12 for the related assumptions. Define the associated sequence of optimal scoring functions $\hat{s}^{(b)} \in \mathscr{S}_0$ obtained at *Step 1* when maximizing the statistic $\widehat{W}_{n',m'}(\phi^{(b)}, \cdot)$. The sequence of test statistics for *Step 2* are therefore related to $\widehat{W}^{(b)}_{n'',m''} = \widehat{W}_{n'',m''}(\phi^{(b)}, \hat{s}^{(b)})$, based on $\mathscr{D}_{n'',m''}$.

The simplest aggregated *R*-statistic is defined by $\bar{W}^B_{n'',m''} = (1/B)\sum_{b\leq B} \widehat{W}^{(b)}_{n'',m''}$. From now on we drop the primes. Notice the difficulty here is related to the unknown null distribution of the statistic $\bar{W}^B_{n,m}$We can estimate its quantile using data-driven procedures such as permutation or bootstrap approaches. However, we would lose the fundamental property of rank statistics of obtaining explicit null distribution. That is why, by following the works of Baraud (2002), we can consider a weighted combination of the statistics by introducing the sequence $(\lambda_b)_{b\leq B}$, such that $\sum_{b\leq B}\lambda_b = 1$ and $\bar{W}^B_{n,m,\lambda} = \sum_{b\leq B}\lambda_b \widehat{W}^{(b)}_{n,m}$. The procedure learns the optimal weights to obtain nonasymptotic guarantees for both errors. Hence, for a given testing level α, $\bar{W}^B_{n,m,\lambda}$ rejects $\mathscr{H}_0$ if there exists one test among the Bs rejecting the null at corrected level $\lambda_b t_\alpha$, where t_α is the least conservative threshold such that the aggregate statistic is of level α. Following Baraud (2002), t_α is defined by

$$t_\alpha := \sup_{t\in(0,\min_{b\leq B}(1/\lambda_b))} \left\{ \mathbb{P}_{\mathscr{H}_1}\left\{ \max_{b\leq B} \left(\widehat{W}^{(b)}_{n,m} - q^{\phi^{(b)}}_{n,m}(\lambda_b t) \right) > 0 \right\} \leq \alpha \right\}, \tag{A.2.5}$$

where we recall that $q^{\phi^{(b)}}_{n,m}$ is the quantile of the null distribution of the statistic $\widehat{W}^{(b)}_{n,m}$. Notice also that the sup in Eq. (A.2.5) is well defined for the considered range of t. We propose an estimator of the optimal threshold by randomly sampling $M' \in \mathbb{N}^*$ samples from the alternative distribution, denoted by $\mathscr{D}^{(i)}_{n,m}$, with $i \leq M'$, as follows

$$\hat{t}_\alpha := \sup_{t\in(0,\min_{b\leq B}(1/\lambda_b))} \left\{ \frac{1}{M'}\sum_{i=1}^{M'} \mathbb{I}\left\{ \max_{b\leq B} \left(\widehat{W}^{(b)}_{n,m}(\mathscr{D}^{(i)}_{n,m}) - q^{\phi^{(b)}}_{n,m}(\lambda_b t) \right) > 0 \right\} \leq \alpha \right\}. \tag{A.2.6}$$

In fact, in comparison to the works of Albert et al. (2021); Baraud (2002) (for independence testing) and Schrab et al. (2021) (for two-sample testing), this method is much simpler as the null distributions for all $b \leq B$ statistics are computable, and for all sizes of samples. This leads to the exact computation of the quantile $q^{\phi^{(b)}}_{n,m}$ at all levels. We therefore propose the use of the statistic

$$\widehat{\Phi}_\alpha(\mathscr{D}_{n,m}, \mathscr{C}) = \mathbb{I}\left\{ \max_{b\leq B} \left(\widehat{W}^{(b)}_{n,m}(\mathscr{D}^{(i)}_{n,m}) - q^{\phi^{(b)}}_{n,m}(\lambda_b \hat{t}_\alpha) \right) > 0 \right\}. \tag{A.2.7}$$

Based on this statistic, data-driven procedures can be developed to estimate $\hat{t}_\alpha$, leading to the aggregated version of the *R*-statistic over class $\mathscr{C}$. We importantly highlight that we can compute the exact null distribution of this procedure for all (n, m) and at all test levels. In the mentioned

literature, results are usually obtained under regularity conditions on the class of density functions, such as Besov spaces or Sobolev balls.

Example 86. *This modeling is particularly adapted for the score-generating function associated to the LocalAUC (Clémençon and Vayatis (2007)), wherein the set of hypotheses is indexed by the parameter* $u_0 \in (0.5, 1)$ *and* $\mathscr{C}$ *is composed of the functions* $\phi_{u_0} : u \mapsto u\mathbb{I}\{u \geq u_0\}$. *In addition, sharp upperbounds of the uniform separation rate, and lower bounds of the minimax uniform separation rate when based on an adaptive approach of the estimation of* (A.2.7) *can be studied.*

A.3 Towards an efficient optimization algorithm for high-dimensional two-sample testing.

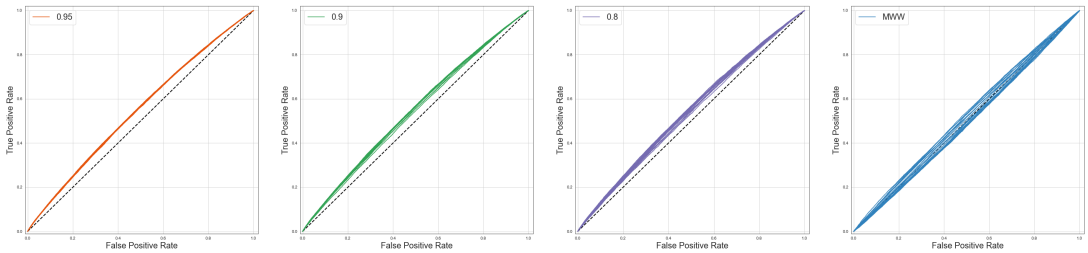
This short section provides some numerical results in the context of the two-sample problem, for a particular type of data structure. Indeed, we explore how to leverage the deterministic algorithm detailed in Chap. 7 when facing very small sample sizes of observations valued in $\mathscr{X} \subset \mathbb{R}^d$, with $d > 1$ greater than n. While providing an interpretable algorithm, it also allows for the exact optimization of the smoothed version of the W_ϕ-ranking performance criterion for multiple choices of score-generating functions ϕ. We sequentially detail the algorithm we perform, the probabilistic models and the parameters for the experiments.

Algorithm, parameters and probabilistic model. We implement the two-stage procedure detailed in the previous section A.2 (Fig. A.1), wherein the algorithm for *Step 1* is chosen to be the Gradient Ascent detailed in Section 7.1, Algorithm 3. Precisely the empirical version of the smoothed counterpart of the W_ϕ-ranking performance criterion is optimized.

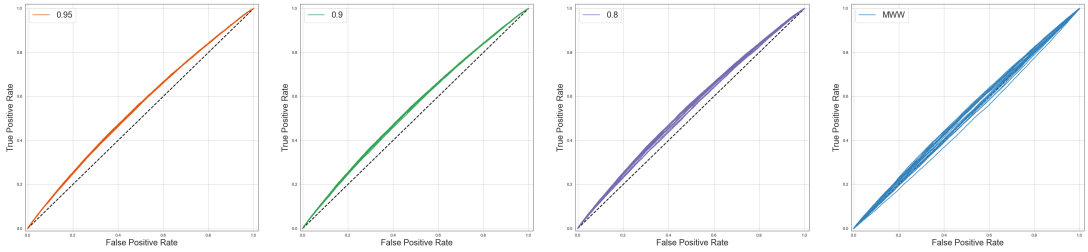
- $\mathscr{S}_0$ is a parametric class, indexed by a parameter space $\Theta \subset \mathbb{R}^d$ with $d \geq 1$ say: $\mathscr{S}_0 = \{s_\theta : \mathscr{X} \to \mathbb{R},\ \theta \in \Theta\}$.
- We implemented the algorithm for various choices of score-generating functions, in order to illustrate the importance of its choice. We considered $\phi_{MWW}(u) = u$ (MWW), and $\phi_{RTB}(u) = \mathrm{SoftPlus}(u - u_0) + u_0\mathrm{Sigmoid}(u - u_0)$, $u_0 \in (0,1)$ (RTB, smoothed version of Clémençon and Vayatis (2007)). As for Section 7.1, the activation functions are defined by: $\mathrm{SoftPlus}(u) = (1/\beta)\log(1+\exp(\beta u))$ and $\mathrm{Sigmoid}(u) = 1/(1+\exp(-\lambda u))$, $\beta, \lambda > 0$ being hyperparameters to fit and control the derivative's slope. We let $u_0 \in \{0.8, 0.9, 0.95\}$.
- We sampled the two-samples according to location model (L2) that we detail below: $\mathbf{X} \sim \mathscr{N}_d(\mu_X, \Sigma)$ and $\mathbf{Y} \sim \mathscr{N}_d(\mu_Y, \Sigma)$ are drawn independently, with: $\mu_Y = 0_d$, $\mu_X = (\varepsilon/\sqrt{d}) \times \mathbf{1}_d$, $\Sigma_{i,j} = \beta^{|i-j|}$, for $i, j \leq d$, $\beta = 0.8$, such that $\Sigma \in S_d^+(\mathbb{R})$, $d \in \{30, 50\}$, and $\varepsilon \in \{0.4, 0.8, 1.5, 3.0, 10.0\}$.
- The parameters for the GA are similar to the ones of Section 7.1, except that we let the algorithm run for $T = 200$ loops.

Results. We gather the results in Fig. A.2 (for $d = 30$), A.3 (for $d = 50$). Some points of the discussion in Section 7.1 remain true for these experiments. Briefly, the present results are very promising as we empirically obtain that the statistics associated to RTB recover the oracle curve (beginning) ROC* with smaller variance compared to MWW. Also, the more proportion u_0 increases and the more the variance of the associated criterion is low. In fact, even if it remains to perform the exact *Step 2*, these results motivate to consider the alternative procedure detailed in Fig. A.1 for such

complex data structures. We highlight that these are preliminary results and we intend to test broader scale of probabilistic models in the future. Importantly this procedure has tractable optimization algorithm and exact homogeneity test statistic that allows for very high-dimensional data analysis.

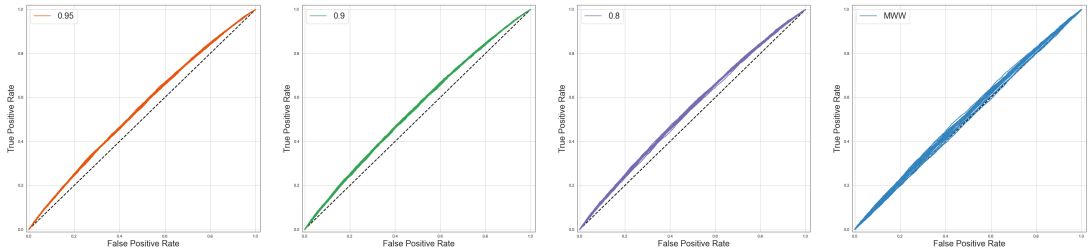


1.a. RTB $u_0 = 0.95$ 1.b. RTB $u_0 = 0.9$ 1.c. RTB $u_0 = 0.8$ 1.d. MWW

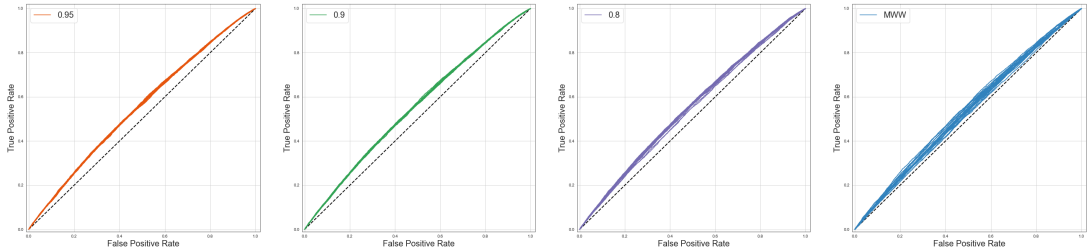


2.a. RTB $u_0 = 0.95$ 2.b. RTB $u_0 = 0.9$ 2.c. RTB $u_0 = 0.8$ 3.d. MWW

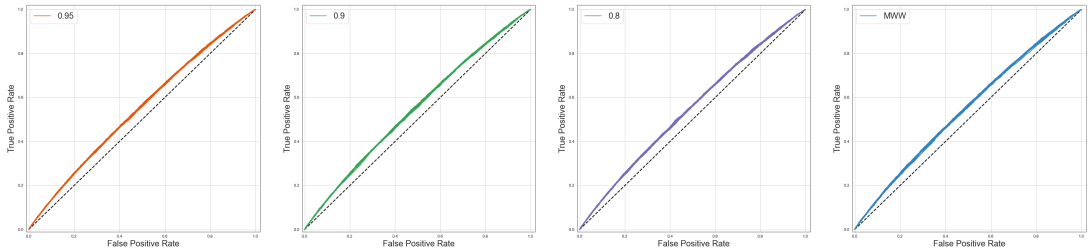
Figure A.2. Empirical ROC curves for (L3) with $\varepsilon = 0.4$ for (1.) and $\varepsilon = 0.8$ for (2.). Maximization of the W_ϕ-criterion with ϕ_{rtb} for figures (a-c) with $u_0 \in \{0.95, 0.9, 0.8\}$ and with ϕ_{mww} for figures (d). Samples are scored with early-stopped GA algorithm's optimal parameter for the class of scoring functions. Hyperparameters: $B = 20$, $T = 200$. Parameters for the training set: $n = m = 24$; $d = 30$; for the testing set: $n = m = 10^5$; $d = 30$.



1.a. RTB $u_0 = 0.95$ 1.b. RTB $u_0 = 0.9$ 1.c. RTB $u_0 = 0.8$ 1.d. MWW

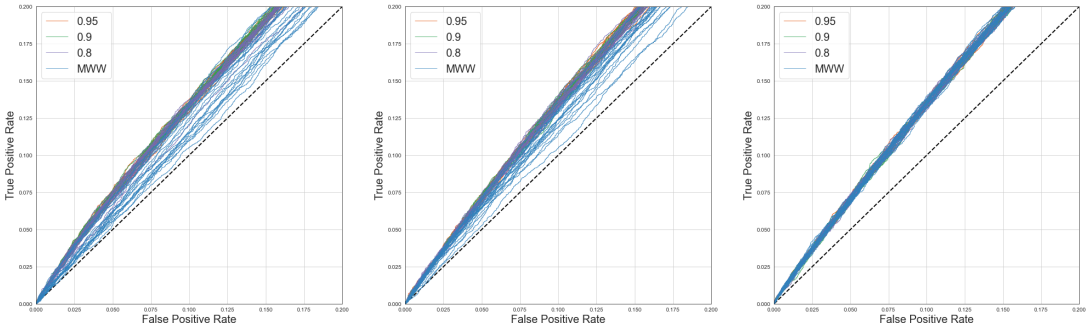


2.a. RTB $u_0 = 0.95$ 2.b. RTB $u_0 = 0.9$ 2.c. RTB $u_0 = 0.8$ 2.d. MWW



3.a. RTB $u_0 = 0.95$ 3.b. RTB $u_0 = 0.9$ 3.c. RTB $u_0 = 0.8$ 3.d. MWW

Figure A.3. Empirical ROC curves for (L3) with $\varepsilon = 1.5$ for (1.), $\varepsilon = 3.0$ for (2.), and $\varepsilon = 10.0$ for (3.). Maximization of the W_ϕ-criterion with ϕ_{rtb} for figures (a-c) with $u_0 \in \{0.95, 0.9, 0.8\}$ and with ϕ_{mww} for figures (d). Samples are scored with early-stopped GA algorithm's optimal parameter for the class of scoring functions. Hyperparameters: $B = 20$, $T = 200$. Parameters for the training set: $n = m = 24$; $d = 50$; for the testing set: $n = m = 10^5$; $d = 30$.



1. (L3), $\varepsilon = 3.0$ 2. (L3), $\varepsilon = 5.0$ 3. (L3), $\varepsilon = 10.0$

Figure A.4. Comparison on the beginning of the empirical ROC curves for the (L3) model and for RTB (orange, green and purple) with $u_0 \in \{0.95, 0.9, 0.8\}$ and MWW (blue). Samples are scored with early-stopped GA algorithm's optimal parameter for the class of scoring functions. Hyperparameters: $B = 20$, $T = 200$. Parameters for the training set: $n = m = 24$; $d = 50$; for the testing set: $n = m = 10^5$; $d = 30$.

B | Univariate framework and state-of-the-art

This section spotlights fundamental topics and methods in univariate statistics for this thesis. First, some definitions and properties of rank statistics (R-statistics) are recalled, the two-sample problem is stated and properties of hypothesis testing are enumerated. A graphical tool is then introduced, namely the Receiver Operating Characteristics (ROC) curve, especially used in biomedicine.

B.1 Univariate rank statistics

Rank statistics are a particular example of permutation statistics, for which they depend on the observations only through their relative *order* when compared to the whole sample.

Introduction to R-statistics. Consider the set of $N \in \mathbb{N}^*$ real observations $z_1, \dots, z_N$. We define their ordered sequence as $z_{(1)}, \dots, z_{(N)}$ such that it is possible to range them in ascending order: $z_{(1)} \leq \dots \leq z_{(N)}$. In fact, when considering a sequence of real and independent random variables (*r.v.*) $Z_1, \dots, Z_N$, we similarly define the *order statistics* as those ranged by ascending order: $Z_{(1)}, \dots, Z_{(N)}$ such that $Z_{(1)} \leq \dots \leq Z_{(N)}$, almost surely. These statistics induce a notion of ranking that takes form of *rank statistics*. The sequence of rank variables associated to the $Z_1, \dots, Z_N$ is defined as $R_1, \dots, R_N$, such that

$$Z_i = Z_{(R_i)}, \quad \text{for all} \quad i \leq N \, .$$

For a given observation, the corresponding rank equals to its position *w.r.t.* the order statistics, when no ties are supposed *i.e.* if the random variables have continuous distribution functions so that the ties occur with probability zero. Under more general assumption, we choose the definition of *upranks* as follows

$$\text{Rank}(Z_i) = \sum_{j \leq N} \mathbb{I}\{Z_j \leq Z_i\}, \quad \text{for all} \quad i \leq N \, .$$

Some basic results are formulated below.

Lemma 87. *(Lemma 13.1, van der Vaart (1998)) Suppose the sequence of independent r.v. $Z_1, \dots, Z_N$, $N \in \mathbb{N}^*$ has common continuous distribution function, then:*

(i) the vectors $\{Z_{(i)}\}_{i \leq N}$ and $\{R_i\}_{i \leq N}$ are independent

(ii) the vector $\{R_i\}_{i \leq N}$ is uniformly distributed over all the $N!$ permutations of $\mathcal{S}_N$

We define the *rank statistics* as functions of the rank variables and consider the particular form of *simple linear rank statistics*, defined by

$$T_N = \sum_{i \leq N} c_{Ni} a_{N,R_{Ni}} \ , \tag{B.1.1}$$

where the subscript N is to specify the size of the considered vectors. The sequences $\{a_{N,i}\}_{i\leq N}$ and $\{c_{Ni}\}_{i\leq N}$ are respectively defined as the *scores* and the *coefficients* of the statistic.

Example 88. (TWO-SAMPLE R-STATISTIC) *Let $n \in \mathbb{N}^*$ and $m = N-n$. Suppose the n first variables constitute the first sample, while the remaining m form the second one, then the coefficients of the linear R-statistic are defined by*

$$(c_{N1}, \dots, c_{NN}) = (\underbrace{1, \dots, 1}_{n \text{ times}}, \underbrace{0, \dots, 0}_{m \text{ times}})$$

and $\bar{c}_N = n/N$, $\sum_{i=1}^N (c_{Ni} - \bar{c}_N)^2 = nm/N$.

The following Lemma is a simple illustration that emphasizes the simplicity of these remarkable statistics.

Lemma 89. *(Lemma 13.1, van der Vaart (1998)) Suppose the sequence of r.v. $Z_1, \dots, Z_N$, $N \in \mathbb{N}^*$ has common continuous distribution function, then*

$$\mathbb{E}[T_N] = N \bar{c}_N \bar{a}_N \quad \text{and} \quad \mathrm{Var}[T_N] = \frac{1}{N-1} \sum_{i=1}^N (c_{Ni} - \bar{c}_N)^2 \sum_{i=1}^N (a_{Ni} - \bar{a}_{N,i})^2 \ , \tag{B.1.2}$$

where $\bar{c}_N = (1/N)\sum_{i\leq N} c_{Ni}$ and $\bar{a}_N = (1/N)\sum_{i\leq N} a_{N,i}$.

In fact the scores $\{a_{N,i}\}_{i\leq N}$ are usually supposed to be generated from a *score-generating function* $\phi : [0,1] \to \mathbb{R}$, for which two specific definitions are commonly used for R-statistics. First, consider a sequence of *i.i.d. r.v.* $U_1, \dots, U_N$ uniformly drawn in $[0,1]$, then for all $i \leq N$,

$$a_{N,i} = \mathbb{E}[\phi(U_{N(i)})] \ , \tag{B.1.3}$$

where $\{U_{N(i)}\}_{i\leq N}$ is the order N-sample of the sequence $U_1, \dots, U_N$. The second definition is

$$a_{N,i} = \phi\left(\frac{i}{N+1}\right) \ . \tag{B.1.4}$$

Under assumptions on ϕ, both definitions are related as $\mathbb{E}[U_{N(i)}] = i/(N+1)$. If the *r.v.* Z is absolute continuous of square-integrable probability density function f and distribution function F, a third representation due to Hajek (1961) and resulting from a projection method, is a tool for asymptotic results such as UMP tests. In particular, the score-generating function in defined as $\phi(i/(N+1)) = \int_{(i-1)/N}^{i/N} -f'(F^{-1}(u))/f(F^{-1}(u))du$, for all $i \leq N$. Interesting examples for the two-sample model (Example 88) are detailed below.

Example 90. (CLASSIC TWO-SAMPLE TESTS) *Two-sample rank statistics applied to hypothesis testing.*

- *The median test is generated by defining the scores with (B.1.4) with $\phi(u) = \mathbb{I}\{(0,1/2])\}(u)$, s.t.*

$$T_{n,m} = \sum_{i=1}^n \mathbb{I}\{R_{Ni} \leq (N+1)/2\} \ .$$

- *The van der Waerden test. Let Φ the cumulative distribution of a standard normal variable, then*

$$T_{n,m} = \sum_{i=1}^{n} \Phi^{-1}(R_{Ni}) \ .$$

- *The Wilcoxon test is generated by defining the scores with (B.1.4) with $\phi(u) = u$, s.t.*

$$T_{n,m} = \sum_{i=1}^{n} R_{Ni} \ .$$

The first step for understanding these statistics is to provide their asymptotic behavior as their structure does not allow for a direct analysis. Classic approaches decompose rank statistics. We consider either of the two definitions (B.1.3), (B.1.4), such that the linearized statistic is considered:

$$\bar{T}_N = N\bar{c}_N\bar{a}_N + \sum_{i=1}^{N}(c_{Ni} - \bar{c}_N)\phi(F(X_i)) \ . \tag{B.1.5}$$

The following theorem proves the asymptotic equivalence of the two statistics: $\bar{T}_N$ and T_N.

Theorem 91. *(Theorem 13.5, van der Vaart (1998)) Consider a i.i.d. sequence $Z_1, \ldots, Z_N$ of continuous distribution function F. Suppose either of the two models:*

(i) the scores are defined by (B.1.3) with ϕ not constant a.e. and s.t. $\int_0^1 \phi^2(u)du < \infty$

(ii) the scores are defined by (B.1.4) with ϕ continuous a.e., nonconstant, s.t. $(1/N)\sum_{i\leq N}\phi^2(i/(N+1)) \to \int_0^1 \phi^2$

Then, the sequences T_N (B.1.1) and $\bar{T}_N$ (B.1.5) are asymptotically equivalent, of same mean and s.t. $\mathrm{Var}[T_N - \bar{T}_N]/\mathrm{Var}[T_N] \to 0$.

Linear *R*-statistics under alternatives. One of the major applications of R-statistics is hypothesis testing where the null hypothesis to test is related to the equality in the underlying distributions, against the alternatives, usually formulated such that each variables is drawn from a different distribution. We consider the generalized formulation, where the sequence of *r.v.* $Z_1, \ldots, Z_N$ is respectively drawn from the continuous distribution functions $F_1, \ldots, F_N$. A straightforward consequence is that the rank variables are no longer uniformly drawn on the set of permutations (Lemma 87-(ii)). The structure of R-statistics is ever more complex under alternatives, formed of a correlated sum of non-*i.i.d. r.v.*. An equivalent statistic of simpler form, at least asymptotically, is thus necessary. A classic method introduced by J. Hájek in his seminal contribution Hájek (1968), corresponds to projecting the statistic onto the space induced by the independent *r.v.*. It is detailed in the sequel for general form of square integrable statistics, obtaining a linearized statistic composed of an average of independent *r.v.*plus an uncorrelated term. Refer also to Chapter 11 in van der Vaart (1998) for further details.

Lemma 92. (HÁJEK PROJECTION, HÁJEK (1968)) *Let a sequence of i.i.d. r.v. $Z_1, \ldots, Z_N$ and $S_N = S_N(Z_1, \ldots, Z_N)$ a real-valued square integrable statistic. The Hájek projection of S_N is defined as*

$$\widehat{S}_N = \sum_{i=1}^{N} \mathbb{E}[S_N \mid Z_i] - (N-1)\mathbb{E}[S_N] \ . \tag{B.1.6}$$

It is the orthogonal projection of the square integrable r.v. S_N onto the subspace of all variables of the form $\sum_{i\leq N} g_i(Z_i)$, for arbitrary measurable functions g_i s.t. $\mathbb{E}[g_N^2(Z_i)] < +\infty$.

Lemma 92 applied to T_N with scores generated by (B.1.4) is important in Chap. 5. Consider the sample $Z_1,\ldots,Z_N$ of independent variables and of respective distribution functions $F_1,\ldots,F_N$, suppose $\phi:[0,1]\to\mathbb{R}$ twice continuously differentiable. Using $\text{Rank}(t)=N\widehat{F}_N(t)$ and $\bar{F}_N=(1/N)\sum_{i\leq N}F_i$, one can consider

$$\bar{T}_N=\sum_{i=1}^{N}c_{Ni}\phi\left(\bar{F}_N(Z_i)\right)\ . \tag{B.1.7}$$

By regularity of ϕ and of the sequence of *c.d.f.* $F_1,\ldots,F_N$, a Taylor expansion of order two around $\bar{F}_N$ of T_N yields

$$T_N=\bar{T}_N+\sum_{i=1}^{N}c_{Ni}\left(\frac{R_{Ni}(Z_i)}{N+1}-\bar{F}_N(Z_i)\right)\phi'\left(\bar{F}_N(Z_i)\right)$$
$$+\sum_{i=1}^{N}\frac{c_{Ni}}{2}\left(\frac{R_{Ni}(Z_i)}{N+1}-\bar{F}_N(Z_i)\right)^2\phi''\left(\bar{F}_N(Z_i)\right)+\mathscr{O}_{\mathbb{P}}(N^{-1})\ .$$

Hájek's lemma 92 applied to the second term and the asymptotic bounding on the second term plus the remainder term of the projection, yields the asymptotic equivalent statistic

$$\widehat{T}_N=\bar{T}_N+\sum_{i=1}^{N}\int_{X_i}^{\infty}\phi'\left(\bar{F}_N(u)\right)d\bar{F}_N^c(u)\ , \tag{B.1.8}$$

where the weighted average equals to $\bar{F}_N^c=(1/N)\sum_{i\leq N}c_{Ni}F_i$. We refer to Section 13.4 of van der Vaart (1998) for all technical details. A first inequality for bounding the variance of the difference of the two statistics is obtained in the following. Hájek (1968) (Theorem 3.1) obtained a weaker bound for more general score-generating functions.

Lemma 93. *(Lemma 13.23, van der Vaart (1998)). Let $\phi:[0,1]\to\mathbb{R}$ twice continuously differentiable. Then, there exists a universal constant $K>0$, such that*

$$\text{Var}(T_N-\widehat{T}_N)\leq\frac{K}{N}\sum_{i=1}^{N}(c_{Ni}-\bar{c}_N)^2\left(\|\phi'\|_\infty^2+\|\phi''\|_\infty^2\right)\ . \tag{B.1.9}$$

Some notes on fundamental results. We formulate the rank central limit theorem under the hypothesis of identical probability distributions and for generic sequence of scores a_N, as defined in Eq. (B.1.1) and (B.1.5). The result is obtained thanks to J. Hájek.

Theorem 94. *(Theorem 4.4, Hajek (1961)) Let $Z_1,\ldots,Z_N$ a i.i.d. sequence of r.v. and define T_N as in Eq. (B.1.1). Suppose $\max_{i\leq N}|c_{Ni}-\bar{c}_N|/C_N\to 0$ and similarly $\max_{i\leq N}|a_{Ni}-\bar{a}_N|/A_N\to 0$, with $C_N=\sum_{i\leq N}(c_{Ni}-\bar{c}_N)^2$ and $A_N=\sum_{i\leq N}(a_{Ni}-\bar{a}_N)^2$. Then,*

$$\frac{T_N-\mathbb{E}[T_N]}{\sqrt{\text{Var}[T_N]}}\longrightarrow W\ , \tag{B.1.10}$$

where W is a standard normal variable, iff for all $\varepsilon>0$,

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{|c_{Ni}-\bar{c}_N|^2|a_{Nj}-\bar{a}_N|}{A_N^2C_N^2}\to 0\ . \tag{B.1.11}$$

Besides the classic asymptotic results, the class of R-statistics challenged the standard parametric tests, where a fundamental question of practitioners and statisticians is which statistics should they use to obtain *better performances*. The answer depends on the probabilistic model and the related asymptotic (relative) efficiency, Bahadur and Pitman efficiencies, (local) power and contiguous alternatives, *etc.*, see van der Vaart (1998). In particular, a major challenge related to rank-based inference is how to choose the optimal score-generating function ϕ. The works of H. Koul, later generalized and gathered in Koul (2002) for the location multivariate model, is a reference. The works initialized by J. Hájek (Hájek (1962, 1968); Hájek and Sidák (1967)) and later by H. L. Koul and J. Jurecková are of high importance for classes of R-statistics and processes, generalized on multiple samples, studying the asymptotic properties under alternatives for very generic forms of linear statistics, see Gutenbrunner and Jurečková (1992); Jurečková et al. (2010).

R-statistics as U-statistics. The two-sample Wilcoxon statistic (Wilcoxon (1945)) is related to two-sample U-statistics: consider two independent samples $X_1,\ldots,X_n$ and $Y_1,\ldots,Y_m$, then the ranksum statistic $W=\sum_{i\leq n}\mathrm{Rank}(X_i)$ can be decomposed by means of U-statistics. If we define $\psi(x,y)=\mathbb{I}\{x>y\}$ for simplicity in the absence of ties, and define $U_{n,m}=\sum_{i\leq n}\sum_{j\leq m}\psi(X_i,Y_j)$, then

$$W_{n,m}=U_{n,m}+n(n+1)/2\,.$$

U is in fact the Mann-Whitney U-statistic (Mann and Whitney (1947)), that is proportional to the generalized U statistic: $U_{n,m}=(nm)^{-1}U$.

B.2 The two-sample problem

Suppose we want to statistically test the efficiency of a medical treatment for which, typically, practitioners would choose two groups with similar characteristics (*e.g.* age, medical history, symptoms). The experimental protocol then would recommend to administrate the drug to be tested to the first group while a placebo would be given to the second one. The problem consists in testing statistically its effect, based on the features/measures collected for both populations after an observational time. We suppose that practitioners know neither the typical behavior of the populations nor how the drug affects the patients.

Formally, this two-sample test is modeled as follows. Given two independent random variables X and Y, valued in the measurable space $\mathscr{X}$, either univariate or multivariate, of continuous distribution functions G and H, the goal is to test, for a fixed level $\alpha\in(0,1)$ the null hypothesis of the equality of distributions:

$$\mathscr{H}_0:\ G=H \text{ against the alternative } \mathscr{H}_1:\ G\neq H\,. \tag{B.2.1}$$

This formulation being quite generic, many problems can be related, such as to goodness-of-fit (GoF) testing, see for instance Darling (1957) (univariate) and Friedman (2004) (multivariate), independence testing (*e.g.* Spearman (1904)) and pairwise testing (*e.g.* Wilcoxon (1945)). The two-sample problem is in practice formulated based on two independent *i.i.d.* sequences $\{X_1,\ldots,X_n\}$ and $\{Y_1,\ldots,Y_m\}$, *resp.* drawn from G and H, with $n,\,m\in\mathbb{N}^*$, such that statistic for testing $\mathscr{H}_0$ is based on the empirical measures

$$\widehat{\mu}_n=\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}\text{ and }\widehat{\nu}_m=\frac{1}{m}\sum_{j=1}^{m}\delta_{Y_j}\,, \tag{B.2.2}$$

or similarly on both empirical *c.d.f.* $\widehat{G}_n(t)=(1/n)\sum_{i=1}^{n}\mathbb{I}\{X_i\leq t\}$ and $\widehat{H}_m(t)=(1/m)\sum_{j=1}^{m}\mathbb{I}\{Y_j\leq t\}$, and many generalizations of univariate two-sample tests rely on related empirical metrics such as quantiles, copulas, data depth functions, *etc.* While there exists a plethora of statistics in this very

generic formulation, the present thesis is restricted to nonparametric testing. Refer to classic books Gibbons and Chakraborti (2011); Lehmann and Romano (2005); Sheskin (2011); van der Vaart (1998) for comprehensive reviews of theory, methodologies and statistics in the field of (nonparametric) hypothesis testing.

Example 95. (LOCATION AND SCALE TESTS) *In (semi)parametric testing, the location/shift problem is formulated, by considering* $G(t) = \mu(t-\theta_1)$, $H(t) = \nu(t-\theta_2)$, *with parameters* $\theta_1,\ \theta_2 \in \mathbb{R}$*:*

$$\mathscr{H}_0 :\ \theta_1 = \theta_2 \quad vs. \quad \mathscr{H}_1 :\ \theta_1 \neq \theta_2\ .$$

It is usually assumed $\mu,\ \nu$ *known and equal. Notice that the simplest version is the test of means for Gaussian distributions, recovering Student's t-test. The scale/shape test can similarly be defined, let* $\theta_i, \sigma_i \in \mathbb{R} \times \mathbb{R}_+^*$, *then* $G(t) = \mu((t-\theta_1)/\sigma_1)$ *and* $H(t) = \nu(((t-\theta_2)/\sigma_2)$, *such that the hypothesis to test are:*

$$\mathscr{H}_0 :\ \theta_1 = \theta_2,\ \sigma_1 = \sigma_2 \quad vs. \quad \mathscr{H}_1 :\ \theta_1 = \theta_2,\ \sigma_1 \neq \sigma_2\ .$$

Some statistical properties. The subsequent properties for both theoretical and practical aspects are relevant and adequate for the characterization of the various test statistics.

(E1) *Type-I error* or the false positive proportion, corresponds to the probability of falsely rejecting $\mathscr{H}_0$.

(P) *Type-II error* or *power*: while constructing a test statistic highlighting the discrepancies of the two distributions may be straightforward, it is by far more complicated to control and maximize its power. It is defined as the probability of rejecting the alternative distribution. Usually, it is proved under fixed alternatives, but some results can be established for sequences of local alternatives, leading to contiguity analysis. This concept was introduced by Cam (1960), later obtaining the famous Le Cam's First and Third lemmas.

(C, AC) *Consistency* and *asymptotic consistency* against any fixed alternative hypothesis, for establishing convergence in probability of the statistic to its expected value or equivalently states the convergence of the power to one, when the number of observations tends to infinity.

(DF) *Exact/asymptotic distribution freeness* with respect to $\mathscr{H}_0$. It is typically a property of univariate rank-based tests.

(DFa) *Essential maximal ancillary* refers to the concept of (DF) but *w.r.t.* the alternative. Ancillarity was introduced by Fisher (1925) and the concept of maximal ancillary later by Basu (1959). Briefly, a statistic is defined as *ancillary* if it is invariant *w.r.t.* transformations of the alternative probabilistic model. It is said maximal if there does not exist another ancillary statistic that can be written as function of it, without being equivalent. For instance, rank statistic satisfy this definition for translation transformations. In particular Basu studied the case where statistics could be modified in sets of probability zero. Refer also to Lehmann and Scholz (1992) for examples.

(U) *Unbiased* test: for a test of size $\alpha \in (0,1)$, the power of the test is lowerbounded by α where also α upperbounds the type-I error, uniformly over all the admissible sets. If there exists a Uniformly Most Powerful (UMP) test, it is unbiased. If the asymptotic distribution of the statistic is Gaussian, under the null and local alternatives, then it is asymptotically locally unbiased.

Below, three classic univariate tests subject to generalizations in the multivariate framework (detailed in Chap. 2), are essential to illustrate some statistical concepts for which the exact tabulation for small sample sizes are available.

Mann-Whitney Wilcoxon (MWW) test. It was introduced in Wilcoxon (1945) as a rank-based statistic $T_{n,m}$, and in Mann and Whitney (1947) as an unbiased U-statistic $U_{n,m}$. The ranks of one sample among the pooled sample are summed up and the famous relation that allows for defining the Mann-Whitney-Wilcoxon statistic is

$$T_{n,m} = nmU_{n,m} + n(n+1)/2 \,. \tag{B.2.3}$$

It is known as the uniformly most powerful test for the location problem and fulfills all the listed properties. In particular, (U) is satisfied against one-sided alternative and if the distributions G and H are symmetric, it is also satisfied against two-sided alternative.

Kolmogorov-Smirnov (KS) test. The KS test was introduced in Darling (1957); Smirnov (1939) and defined by

$$D_{n,m} = \sup_{z \in \mathscr{Z}} |\widehat{G}_n(z) - \widehat{H}_m(z)| \,. \tag{B.2.4}$$

Smirnov (1939) proved the asymptotic distribution under the alternative to be independent of the data, whereas it remains unknown under the null. The associated metric on the space of distribution functions leads to, *e.g.*, goodness-of-fit testing, but also generalized to empirical processes, see van der Vaart and Wellner (1996). Also, applying the Glivenko-Cantelli theorem or Donsker theorem, yield to consistent test with explicit rate of convergence against any fixed tw-sided alternative, where $D_{n,m}$ is expected to be large under $\mathscr{H}_1$ and to tend to zero otherwise. Numerical experiments show that while performing very well for the location and the scale models, it lacks power for others test. Notice that a related statistic, also known as the Cramér-von Mises, computes the square of the difference between the empirical *c.d.f.*

Wald–Wolfowitz runs (WWr) test. The WWr statistic, introduced in Wald and Wolfowitz (1940), counts the number of runs obtained in the pooled sample. Associate a symbol $+$ to the Xs sample and $-$ to the Ys sample. A *run* is a consecutive sequence of maximal non-empty segment of the sequence consisting of adjacent identical elements, for example

$$++---+-----++++++--+$$

equals to 10 where 5 for $+$ and 5 for $-$. The statistic considered is defined

$$W = \frac{R - 2mn/N - 1}{\sqrt{2mn(2mn-N)/(N^2(N-1))}} \,, \tag{B.2.5}$$

where R denotes the total number of runs. It is shown to be less powerful than the KS for the location problem but more for the scale test. While this method is studied through an extensive literature in the 1940-50s by means of combinatorial approaches Wolfowitz (1943), it has been adapted to Markovian analysis and gathered in Fu (1996).

Remark 14. (TWO-SAMPLE TESTING WITH R-STATISTICS) *The alternative can be formulated as a one-sided hypothesis, yielding*

$$\mathscr{H}_0 : \, G = H \text{ against the alternative } \mathscr{H}_1 : \, G \leq_{sto} H \,, \tag{B.2.6}$$

where $\mathscr{H}_1$ corresponds to H is stochastically larger[1] than G. It corresponds to the very general framework wherein for example, when testing the effect of a treatment, one can suppose that it cannot

[1] Given two distribution functions $H(dt)$ and $G(dt)$ on $\mathbb{R} \cup \{+\infty\}$, it is said that $G(dt)$ is *stochastically larger* than $H(dt)$ *iff* for any $t \in \mathbb{R}$, we have $G(t) \leq H(t)$. We then write: $H \leq_{sto} G$. Classically, a necessary and sufficient condition for G to be stochastically larger than H is the existence of a coupling $(\mathbf{X}, \mathbf{Y})$ of (G,H), *i.e.* a pair of random variables defined on the same probability space with first and second marginals equal to H and G respectively, such that $\mathbf{X} \leq \mathbf{Y}$ with probability one.

be modeled by a fixed parameter (e.g. location model $G(t) = H(t-\theta)$) but rather suppose that the size of the effect depends on the value of the observation (e.g. θ depends on t). For instance, MWW is UMP when choosing H as the logistic distribution $H(t) = 1/(1+e^{-x})$ related to the location model, see Lehmann and Romano (2005), Chapter 6.9.

Comparing statistical tests to MWW. Historically, the introductory articles on specific R-statistics suggested their interest based on empirical results. Their experiments were confirmed by the famous article of Hodges and Lehmann (1956) that proved that the asymptotic relative efficiency (ARE) for the rank-based Wilcoxon statistic compared to the Student t-test for various null distributions, never goes bellow 0.864. This article had a great impact on the community as mentioned by E.L. Lehmann himself: *"this paper was influential in the sense that it dispelled the belief that while nonpammetric [rank-based] techniques are very convenient because you don't have many assumptions, they have so little power that they are no good"*, interview in *Statistical Science* in 1984, see also Hallin and Tribel (2000).

B.3 Univariate ROC analysis

This section introduces a gold standard tool, known as the Receiver Operating Characteristics (ROC) curve. The ROC curve is a graphical tool formally introduced by Egan (1975) for quantifying the difference between two classes at different thresholds, when based on parametric probability laws. However it was initially motivated by signal detection theory, it became essential to decision making and classifier/model selection. It offers a visual tool gathering performances from different devices, easy to understand and to interpret. ROC curves are used in both datamining and biomedicine, for which a series of references can be found in Krzanowski and Hand (2009) for its properties, and in Fawcett (2006); Park et al. (2004); Swets (1988) for applications.

Two-sample ROC curve as a P-P plot. The ROC curve is a reference tool to describe the dissimilarity between two univariate probability distributions G and H. This functional criterion denoted by $\mathrm{ROC}_{\mathrm{H,G}}$, can be defined as the parametrized curve in $[0,1]^2$

$$t \in \mathbb{R} \mapsto \left(\underbrace{1-H(t)}_{\text{False Positive Rate}} \ , \ \underbrace{1-G(t)}_{\text{True Positive Rate}} \right) . \tag{B.3.1}$$

It directly plots the variation of the *True Positive Rate* (TPR) depending on the *False Positive Rate* (FPR) for all possible threshold levels. TPR, also known as the sensitivity, is therefore plotted against (1- specificity), and corresponds to a Probability-Probability (P-P) plot. The graph is nterpretable by non-experts, justifying its standardization in applied fields.

By convention, we connect possible jumps by line segments to ensure that the resulting curve is continuous. The ROC curve related to the pair of *d.f.* (H,G) is the graph of a càd-làg (*i.e.* right-continuous and left-limited) nondecreasing mapping valued in $[0,1]$, defined by

$$\alpha \in (0,1) \mapsto 1 - G \circ H^{-1}(1-\alpha) \ , \tag{B.3.2}$$

at points α such that $H \circ H^{-1}(1-\alpha) = 1-\alpha$. By denoting $\mathscr{Z}_H$ (*resp.* $\mathscr{Z}_G$) the support of H (*resp.* G), that the ROC curve connects the point $(0, 1-G(\mathscr{Z}_H))$ to $(H(\mathscr{Z}_G), 1)$ in the unit square $[0,1]^2$. In the following, we suppose that there are no atoms of the *c.d.f.* , instead of restricting $\mathscr{Z}$ to $\mathscr{Z}_G$. The curve $\alpha \in (0,1) \mapsto \mathrm{ROC}_{\mathrm{G,H}}(\alpha)$ is the image of $\alpha \in (0,1) \mapsto \mathrm{ROC}_{\mathrm{H,G}}(\alpha)$ by the reflection with the main diagonal of the Euclidean plane as axis. Proposition below gathers fundamental properties.

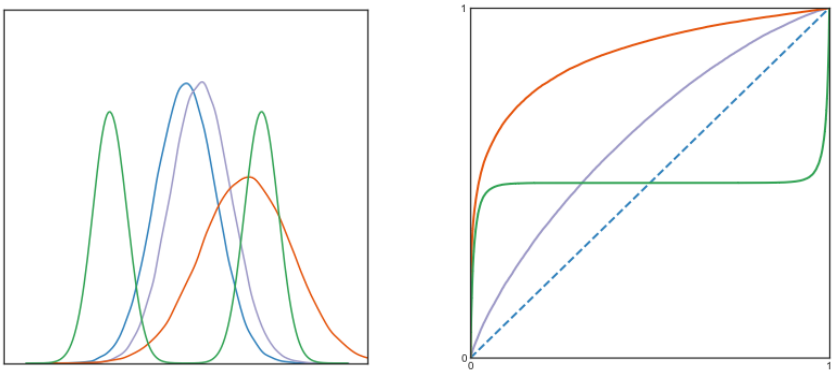
Proposition 96. *Suppose G and H two probability functions. The following assertions hold true.*

(i) $G = H$ *iff* $\mathrm{ROC}_{\mathrm{H,G}}(\alpha) = \alpha$, *for all* $\alpha \in [0,1]$.

(ii) $H \leq_{sto} G$ *iff* $\mathrm{ROC}_{\mathrm{H,G}}(\alpha) \geq \alpha$, *for all* $\alpha \in [0,1]$.

(iii) $\mathrm{ROC}_{\mathrm{H,G}}$ *coincides with the left upper corner of the unit square iff the essential supremum of the distribution H is smaller than the essential infimum of the distribution G.*

Hence, the concept of ROC curve offers a visual tool to examine the differences between two distributions in a pivotal manner, as illustrated in Fig. B.1.



a. Probability distributions b. ROC curves

Figure B.1. Examples of pairs of distributions and their related ROC curves. The distribution H is represented in blue and three examples of G distributions are in purple, orange and green, like the associated ROC curves.

A univariate summary. Another advantage of the ROC curve lies in the probabilistic interpretation the summary criterion, referred to as the Area Under the ROC Curve (AUC in short). It indicates the average dissimilarity of two variables overall the different 'thresholds' discriminating them and is defined by

$$\mathrm{AUC}_{\mathrm{H,G}} := \int_0^1 \mathrm{ROC}_{\mathrm{H,G}}(\alpha)\mathrm{d}\alpha = \mathbb{P}\{\mathrm{Y} < \mathrm{X}\} + \frac{1}{2}\mathbb{P}\{\mathrm{X} = \mathrm{Y}\}, \tag{B.3.3}$$

where (X,Y) denotes a pair of independent *r.v.* with respective marginal distributions H and G. A more sensitive criterion is defined as the partial AUC (pAUC), that averages over a prespecified range of the FPR, see *e.g.* McClish (1989). For instance, the simplest definition is by integrating on the half line of the type $]-\infty, \omega]$, where ω thresholds the FPR. The pAUC thus averages the values lying at the *beginning* of the ROC curve, interpreted as the *best* instances. These summaries can lead to more complex statistics, difficult to optimize as they may not be convex. However, the ethical and economical interpretations of both TPR and FPR can need for such tailored criteria. They find a particular interest for biomedical applications as it is possible to constraint prespecified ranges for either TPR or FPR. We refer to Carrington et al. (2020); Walter (2005); Yang et al. (2019).

C | Additional content

C.1 Some facts on scientific research based on statistics

Testing statistical hypothesis became an essential practice in applied research since the 1950s, including in biomedicine, pharmacology, economics, law, engineering, *etc.*, as it is part of every comparison of real measures/observations regarding the quantification of a phenomenon. Usually, it accompanies/supports protocols that are modeled to compare two independent samples/groups/populations, where a null hypothesis representing the absence of phenomenon ($\mathscr{H}_0$) is statistically tested, against an alternative hypothesis ($\mathscr{H}_1$). A *significant* conclusion can drawn if $\mathscr{H}_0$ is *rejected* based on a small p-value, defined as the probability that the actual data-driven event occurred under the null hypothesis. However simple to state, hypothesis testing is at the heart of a strong debate in the scientific community as it drives many scientific findings. For example, the free biomedical archive *PubMed Central*, registered more than 524 773 articles referring to p-values only for the past five years and more than 135 165 in the last year.

In spite of this attractiveness, p-values and more generally hypothesis testing are poorly and imprecisely used. While intending to provide results that should be reliable and reproducible, the findings are questionable as the methodologies are not rigorous and lack of in-depth analysis. In this sense, researches raised their concerns by publishing reviews that tackle statistical practices since the 1980s, see *e.g.* Bland and Altman (1988); Glantz (1980) and ever more recently, with the acceleration of the capacity of data collection and analysis and the ability to openly share articles in online databases. Some of the striking facts are listed in the following regarding the (bio)medical community. First, reproducibility calls scientific practices into question. Indeed, Baker and Penny (2016) reported in *Nature*, that 90% of researchers surveyed believe in a 'reproducibility crisis'. An online survey in 2014 among 900 members of the American Society for Cell Biology reports only 17.18% disagree that the lack of rigorous statistical analysis plays a role in the impossibility to replicate published results (see `go.nature.com/kbzs2b`), while 71.54% reported being unable to replicate a published result. Begley and Ellis (2012) calls for levelling up the reliability of preclinical studies, as it was estimated based on data from GoogleScholar up to May 2011, that Preclinical research generates more secondary publications if nonreproducible than if reproducible.

Besides, a real challenge is raised regarding the methods and the quality of the published analysis. Hill et al. (1997) evaluates approximately 90% of the articles lack of discussion regarding the statistical hypothesis studied. Based on surveys from 1987 to 2008, Fanelli (2009) reports that 33.7% of studies in a pool of articles used dubious research practices. Thiese et al. (2015) points out abuses in analytical plans *w.r.t.* the statistical design, the lack of transparency and errors in the presentation/description of the data, the statistical method applied, *etc.* More generally, the works of J. Ioannidis question why the majority of scientific findings are false, see *e.g.* Ioannidis (2005). Unfortunately, the COVID-19 pandemic has intensified these facts. For instance, the top medical journal *The Lancet* published a highly shared article on the hydroxychloroquine, that was based on data from a huge amount of hospitals and was proved to be false.

In response to these warnings, highly specialized scientists published statistical guidelines and recommendations for constructing analytical plans and protocols, see for instance Benjamin et al. (2018); Lang and Altman (2015); Moher D. (2001); Perneger (1998b). Moreover, online journals and projects, such as IPOL (https://www.ipol.im), Reproducibility Project: Psychology (RP:P) (https://osf.io/ezcuj/), EQUATOR (www.equator-network.org), aim to enhance these *good* practices. While reviews raise awareness among the scientific community, some question the essence of hypothesis testing. Fortunately, some months ago, the culminant communication lead by the Editor in Chief of the journal *The Annals of Applied Statistics*, see Kafadar (2021), gathered a Task Force to clarify the importance of the use of statistical hypothesis and the role of p-values to replicability, see Benjamini et al. (2021).

C.2 Introduction (en français)

C.2.1 Contexte et motivations

Les tests statistiques non paramétriques de comparaison de deux échantillons en grande dimension. Dans sa formulation statistique la plus générale, le problème à deux échantillons teste l'égalité de deux distributions de probabilité inconnues à un niveau de risque fixé, où deux échantillons aléatoires *i.i.d.* indépendants $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_m$ sont conidérés, évalués sur le (même) espace mesurable $\mathscr{X}$, par exemple de $\mathbb{R}^d$, $d \geq 2$. Bien qu'il existe une vaste littérature pour le cas univarié (voir Lehmann and Romano (2005)), ce problème reste un sujet de recherche pour les cadres multivarié et non paramétrique. En effet, la capacité croissante à collecter de nombreuses données, voire massives, de structures variées et éventuellement biaisées dû au processus de collecte, a fortement défié les modélisations classiques, voir *e.g.* Wang et al. (2019). De tels types de données sont notamment analysées dans des domaines appliqués comme en biomédecine (*e.g.* essais cliniques, génomique), en marketing (*e.g.* tests A/B, systèmes de recommandation), en économie, *etc.* Les méthodes récentes développées en grande dimension reposent généralement sur des distance statistiques. Ces dernières sont estimées grâce à des versions empiriques des mesures de probabilité sous-jacentes, tel que plus la distance diminue, plus les deux échantillons peuvent être qualifiés comme homogènes, voir Biau and Gyorfi (2005); Gretton et al. (2012a). Malheureusement, ces formulations dépendent souvent de la définition intrinsèque de la métrique choisie, et de la représentation ambiante des observations aléatoires. De plus, des propriétés statistiques importantes peuvent manquer, concernant par exemple : le contrôle non asymptotique des erreurs de type I et/ou de type II, le calcul exacte de la distribution nulle, ou même la stabilité par rapport à la dimension de l'espace $\mathscr{X}$ (*e.g.* de d).

Statistiques de rang et méthodes d'apprentissage d'ordonencement. En réponse à l' "hypothèse gaussienne" traditionnelle, les statistiques de rang ont été introduites par Spearman dans Spearman (1904), définissant la statistique de test rhô. En effet, les observations étant uniquement considérées *via* leur ordre relatif, les R-statistiques *"réduisent les "erreurs accidentelles""* (page 81, Spearman (1904)). Elles ont, par la suite, gagné en popularité grâce à leur simplicité, leur rapidité de calcul et en tant que cas particulier des statistiques de permutation. Concernant les tests de comparaison à deux échantillons, les statistiques de rang sont très compétitives lorsque les conditions considérées pour les distributions de probabilité sous-jacentes sont faibles, voir *e.g.* Chernoff and Savage (1958); Hodges and Lehmann (1956). En particulier, elles permettent d'obtenir la distribution explicite sous l'hypothèse nulle (égalité des distributions), tout en garantissant une puissance élevée dans le cadre univarié, voir Chap. 15 dans van der Vaart (1998). La version la plus simple,

connue sous le nom de statistique de test *ranksum* ou Mann-Whitney-Wilcoxon (Mann and Whitney (1947); Wilcoxon (1945)), est célèbre pour être asymptotiquement uniformément plus puissante (UPP) pour le problème de localisation à taille de test fixe et sous distributions logistiques, voir Ex.15.15 dans van der Vaart (1998). Cependant, la définition des statistiques de rang est loin d'être simple en grande dimension, dû au manque de relation d'ordre naturel pour les données multivariées. La littérature s'appuie généralement sur la profondeur ou les rangs spatiaux, qui dépendent fortement de leur définition intrinsèque et sont principalement conçus pour des tests (paramétriques) particuliers, voir *e.g.* Chakraborty and Chaudhuri (2017); Hallin and Paindaveine (2008).

Dans un contexte différent, l'ordonnancement d'observations est devenu fondamental dans de nombreux problèmes d'analyse de données au cours des dernières décennies, *e.g.*, en recherche d'information et en biologie computationnelle. Elles sont définies comme méthodes d'apprentissage d'ordonnancement (*learning-to-rank*), c'est-à-dire apprenant un ordre à partir d'un ensemble d'observations en fonction de leur pertinence/importance/préférence, pour prédire l'ordre de tout "nouvel" échantillon de données. En particulier, l'approche la plus simple pour deux échantillons est connue pour être intimement liée à la statistique Mann-Whitney-Wilcoxon *via* l'analyse *Receiver Operating Characteristic* (ROC), voir Clémençon and Vayatis (2009b). Une série de contributions fondamentales reliant les approches d'apprentissage d'ordonnancement et l'analyse ROC sont, par exemple, Agarwal et al. (2005); Cortes and Mohri (2004). Cependant, et à notre connaissance, seuls Clémençon et al. (2008, 2009) ont théoriquement exploité leur relation aux classes de statistiques linaires de rang. Cette dernière direction motive ce travail et a idéalement pour but de proposer des procédures algorithmiques faciles à utiliser, interprétables et traçable.

Vers des guaranties non asymptotiques. Plus généralement, les structures de données complexes acquéries sous des conditions éventuellement biaisées nécessitent une modélisation statistique non paramétrique et multivariée. Concernant les tests d'hypothèses, les extensions multivariées des statistiques de test univariées classiques présentent une analyse non paramétrique limitée. Valiant (1984) a introduit la théorie *probably approximately correct* (PAC), fournissant un cadre quantifiant la difficulté d'un problème d'analyse de données. Brièvement, en considérant un espace de probabilité $(\Omega, \mathscr{A}, \mathbb{P})$, les bornes PAC définissent formellement le contrôle d'un événement A avec une certaine probabilité, comme suit :

$$\textit{A probabilité fixée } \delta > 0\textit{, pour tout element } \omega \in A(\delta) \textit{ tel que l'évènement } A(\delta) \subset \Omega\textit{,}$$
$$\textit{vérifie } \mathbb{P}\{A(\delta)\} \geq 1-\delta.$$

En fait, les bornes de concentration peuvent être obtenues pour un estimateur Z_N basé sur un échantillon de taille $N \in \mathbb{N}^*$, en trouvant le seuil $t_{\delta,N} > 0$ tel que $\mathbb{P}\{Z_N < t_{\delta,N}\} \geq 1-\delta$, où $A(\delta) = \{Z_N < t_{\delta,N}\}$. Avec probabilité $1-\delta$, la variable aléatoire ne dépasse pas un certain seuil, interprété par la même comme intervalle de confiance (non asymptotique) de l'estimateur. Ce type de bornes est particulièrement utilisé en théorie de l'apprentissage statistique pour étudier les fluctuations aléatoires du risque empirique étant donné un modèle.

Application biomédicale : comparaison de données posturographiques. Cette thèse est motivée par un projet biomédical portant sur la quantification du comportement humain développé dans une équipe de recherche interdisciplinaire[1]. En particulier, dans le cadre de la médecine personnalisée et de la prévention des personnes âgées, une équipe réunissant des mathématiciens, des (bio)statisticiens

[1]Le Centre Borelli est un laboratoire de recherche issu de la fusion récente d'un laboratoire de mathématiques appliquées (CMLA, Ecole Normale Supérieure Paris-Saclay) et d'un laboratoire de neurosciences (COGNAC-G, Université Paris Descartes), où se développent de multiples projets interdisciplinaires.

et des cliniciens de différentes spécialités, étudie le contrôle postural des populations cliniques. Le but est de pouvoir recueillir l'évolution du contrôle postural au travers des suivis cliniques pour en détecter une éventuelle détérioration. Plus précisément, la progression de la fragilité chez les patients agés et parkinsoniens est au cœur du projet. Cette population est sujette à l'instabilité posturale, impliquant des chutes possibles à des âges pour lesquels les interventions chirurgicales sont découragées. Une façon de mesurer le contrôle postural consiste à utiliser des plateformes sensorimotrices enregistrant, pendant un court intervalle de temps, la variation temporelle du déplacement du *Centre de pression* (CoP) (statokinésigramme) du patient. Le protocole expérimental est illustré dans la Figure C.1.

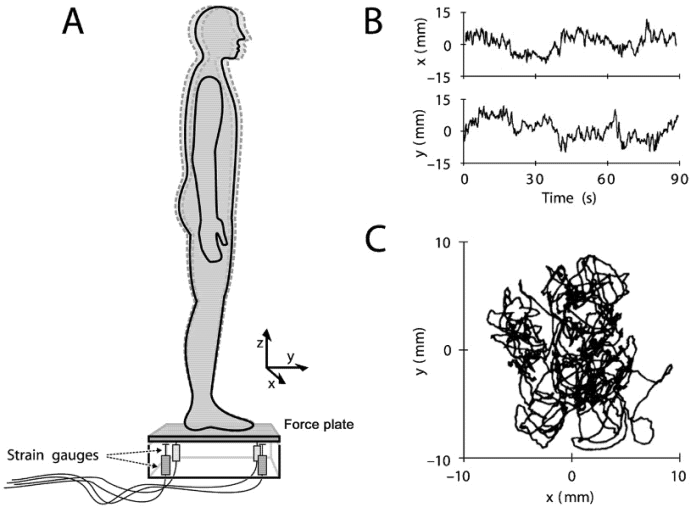


FIGURE C.1. Illustration du protocole d'acquisition des statokinésigrammes. Le patient reste immobile en (A) pour mesurer la trajectoire bi-dimensionnelles du centre de pression par la plateforme de force. Les deux séries temporelles des axes médio-latéral et antéro-postérieur sont resp. représentés sur l'axe x et y en (B). Exemple de statokinésigramme en (C). Source : Chen et al. (2021).

Dans ce contexte, un problème typique et important consiste à comparer des patients au contrôle postural fragile, désignés comme *Chuteur*, à une population 'témoin' choisie, désignée comme *Non Chuteur*. Afin de mieux comprendre la difficulté, la Figure C.2 rassemble des statokinésigrammes mesurés à partir de ces deux populations, pour lesquelles la distinction visuelle entre les paires de patients (a *vs.* b, et c *vs.* d) est loin d'être simple. Les mesures sont de structure complexe (*e.g.* caractéristiques multiples, nature fonctionnelle, cohortes petites/déséquilibrées), pour lesquelles des informations/caractéristiques supplémentaires sur les patients peuvent y être ajoutées (*e.g.* comorbidité, âge). En fait, après un prétraitement adéquat, de nombreuses caractéristiques des statokinésigrammes obtenus peuvent être collectées et analysées, voir Quijoux et al. (2021). Cependant, il existe de fortes contraintes liées à ce type de données pour utiliser des approches traditionnelles de test à deux échantillons. Les praticiens sont confrontés soit à une approche difficilement implémentable, soit à des modèles univariés et paramétriques inadéquats. Nous avons exploré ces approches typiques dans Bargiotas et al. (2021) pour analyser le contrôle postural, et plus généralement mis en évidence quelques faits scientifiques globaux relevés par la communauté scientifique à l'usage des statistiques en Annexe Chap. C.1. Par exemple, pour comparer des observations multivariées, plusieurs procédures de test sont généralement associées à de simples corrections permettant le contrôle de l'erreur de type I voir *e.g.* Hochberg (1988); Hommel (1988). Nous avons comparé leur capacité à discriminer deux populations (*Chuteur*/*Non Chuteur*) à deux méthodes multivariées : la *Maximum Mean Disrepancy* (MMD), voir Gretton et al. (2007), et un algorithme adapté aux données se basant

sur la généralisation des statistiques de rang de Clémençon et al. (2009). Les procédures classiques ne rejettent pas l'hypothèse nulle, *i.e.* , concluent que les deux populations sont tirées de la même distribution. Alors que les deux méthodes multivariées concluent à une différence significative, avec des p-valeurs très faibles. Nous nous référons au chapitre 9 pour les résultats détaillés, aux tableaux 9.3 et 9.4.

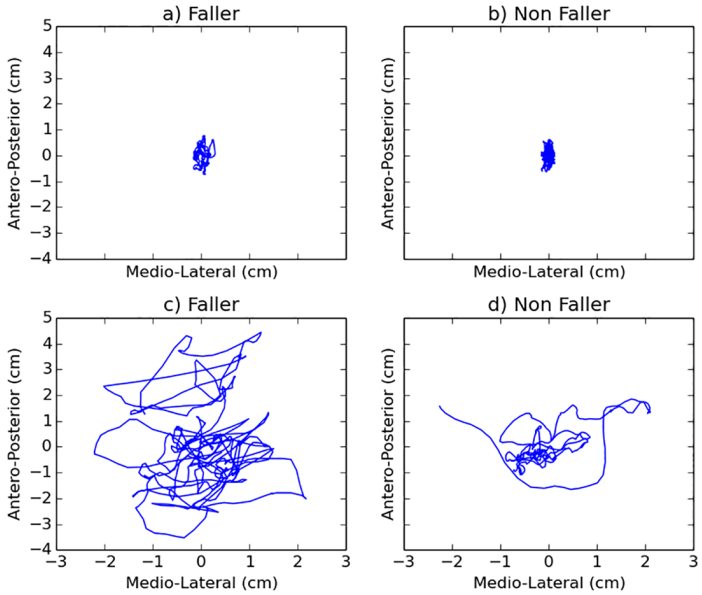


FIGURE C.2. Illustration de statokinésigrammes pour patients *Chuteur* (a,c) et *Non Chhuteur* (b,d) dans l'espace médio-latéral/antéro-postérieur. Source : Audiffren et al. (2016).

C.2.2 Introduction à trois problèmes statistiques fondamentaux à deux échantillons

Tests statistiques de comparaison de deux populations en grande dimension

Cette section introduit le problème à deux échantillons dans le cadre multivarié et non paramétrique. Nous nous référons à la section annexe B.2 pour sa formulation univariée avec un rappel sur les propriétés classiques et les statistiques.

Considérons deux variables aléatoires indépendantes $\mathbf{X}$ et $\mathbf{Y}$, définies sur un espace de probabilité à valeurs dans le (même) espace mesurable multivarié $\mathscr{X}$, de fonctions de répartition continue inconnues G et H. A niveau $\alpha \in (0,1)$ fixé, le problème à deux échantillons correspond au teste de comparaison des deux hypothèses ci-dessous :

$$\mathscr{H}_0 : \; G = H \text{ contre l'alternative } \mathscr{H}_1 : \; G \neq H \; . \tag{C.2.1}$$

Aussi connu sous le nom de teste d'homogénéité, de nombreux problèmes classiques peuvent y être associés. Voir Darling (1957) pour le test d'ajustement (*goodness-of-fit*) et Friedman (2004) pour le cas multivarié, Spearman (1904) pour le test d'indépendance, et enfin Wilcoxon (1945) pour les tests appariés. En pratique, et en particulier pour les variables non paramétriques, nous considérons des copies indépendantes de la *v.a.* étudiée, étant donnée les (classes de) distributions sous-jacentes inconnues. Soient $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ et $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$, avec $n, m \in \mathbb{N}^*$, deux échantillons indépendants et *i.i.d.* tirés suivant G et H, et à valeurs dans le (même) espace mesurable $\mathscr{X}$. Les statistiques univariées non paramétriques, telles que celle de Kolmogorov-Smirnov (Smirnov (1939)), reposent sur des estimations empiriques des distributions sous-jacentes ou des (pseudo)-métriques associées,

voir la section annexe B.2. L'hypothèse nulle $\mathscr{H}_0$ est rejetée si l'on obtient de 'grandes' valeurs de ces statistiques, c'est-à-dire, dans le cas de 'grands écarts' des deux échantillons aléatoires. Pour les observations multivariées, leur distribution empirique sont, par exemple,

$$\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbf{X}_i} \text{ and } \widehat{\nu}_m = \frac{1}{m}\sum_{j=1}^{m} \delta_{\mathbf{Y}_j}\ , \tag{C.2.2}$$

où δ_x est la masse de Dirac en tout point x, ou des versions empiriques des fonctions de répartition, quantiles, copules, profondeurs, *etc.* Des (pseudo-)metriques aléatoires classiques mesurant la dissimilarité entre deux distributions de probabilité sont : la distance χ^2 , la divergence de Kullback-Leibler, la distance d'Hellinger, la distance de Kolmogorov-Smirnov. Se référer à Rachev (1991) pour une revue complète.

Une formulation classique dans la littérature non-paramétrique est connue sous le nom de tests minimax, où l'alternative correspond à la séparation des distributions sous-jacentes au sens d'une métrique, voir *e.g.*, Lam-Weil et al. (2022) étudiant le taux de séparation minimax local défini par la norme L_1 et appliqué à des distributions discrètes, Carpentier et al. (2018) utilisant la norme L_2 en régression linéaire creuse. Nous nous référons en particulier à Albert et al. (2021); Berrett et al. (2021) pour les tests d'indépendance, et à Baraud (2002); Ingster and Suslina (2003, 2000); Lepski and Spokoiny (1999) pour les tests d'ajustement. Enfin, un problème connexe dans la littérature informatique fait référence au problème à deux échantillons en tant que test de propriété, voir par exemple Goldreich et al. (1998); Rubinfeld and Sudan (1996). L'exemple ci-dessous formule un test statistique classique connu sous le nom de test de localisation.

Example 97. (TEST DE LOCALISATION DANS $\mathbb{R}^d$) *En statistique (semi-)paramétrique, soient P_1, $P_2 \in \mathscr{P}$ un modèle probabiliste, tel que $G(t) = P_1(t-\theta_1)$, $H(t) = P_2(t-\theta_2)$, de paramètres $\theta_1,\ \theta_2 \in \mathbb{R}^d$, où $d \in \mathbb{N}^*$, le problème de localisation s'écrit :*

$$\mathscr{H}_0 :\ \theta_1 = \theta_2 \quad vs. \quad \mathscr{H}_1 :\ \theta_1 \neq \theta_2\ .$$

La forme la plus simple est souvent présentée lorsque P_1, P_2 sont supposées connues et égales. Nous retrouvons le test T^2 d'Hotelling's pour l'égalité des moyennes entre distributions gaussiennes.

Alors que des statistiques peuvent être construites pour un modèle probabiliste particulier, *e.g.* modèles gaussiens, elliptiques, ce manuscrit se concentre sur des formulations non paramétriques pour lesquelles l'obtention de garanties statistiques est possible. Plus précisément, nous nous intéressons à la consistance (asymptotique), au contrôle (asymptotique) des deux erreurs statistiques (type-I et type-II), à l'indépendance de la distribution nulle des statistiques de test par rapport au modèle sous-jacent, à l'indépendance des statistiques de test aux transformations du modèle sous l'alternative (également connu sous le nom de statistiques ancillaires Fisher (1925)), la caractéristique non-biaisée de la statistique de test. Voir la section annexe B.2 pour plus de détails et définitions. Se référer aux livres classiques Gibbons and Chakraborti (2011); Lehmann and Romano (2005); Sheskin (2011); van der Vaart (1998) pour des revues complètes de la théorie, des méthodologies et des statistiques dans le domaine des tests d'hypothèses (non paramétriques).

Processus de rang

Cette section introduit la définition univariée choisie des statistiques de rang linéaire à deux échantillons. Se référer à la section annexe B.1 pour une introduction détaillée dans le cadre univarié et en particulier pour les méthodes classiques (*e.g.* projection de Hájek, Hájek (1968)) et les propriétés fondamentales (asymptotiques) sous les hypothèses nulle et alternative.

Historiquement, les statistiques de rang étaient intéressantes grâce à leur simplicité et à la rapidité de calcul pour des échantillons relativement petits, en commençant formellement par le test rhô de Spearman (Spearman (1904)) et plus tard avec le test à deux échantillons de Wilcoxon (Wilcoxon (1945)). Spearman a motivé sa statistique comme une réponse à l'"hypothèse gaussienne" traditionnelle, car les méthodes d'ordonnancement *"réduisent les "erreurs accidentelles""* (page 81, Spearman (1904)), par rapport à celles basées sur la valeur des observations. En effet, les extrêmes, c'est-à-dire, les observations éloignées du comportement 'moyen', ne 'pèsent' pas plus dans le calcul de la statistique. Au contraire, ces observations rares affectent les statistiques qui tiennent compte de leurs valeurs.

Soient deux *v.a.* indépendantes X, Y respectivement tirées suivant G, H et à valeurs dans $\mathscr{X} \subset \mathbb{R}$. Nous considérons les deux échantillons comme suit. Soient $X_1, \ldots, X_n$, avec $n \in \mathbb{N}^*$, suite d'observations *i.i.d.* tirées suivant G, et $Y_1, \ldots, Y_m$, avec $m \in \mathbb{N}^*$, *i.i.d.* tirées suivant H, telles que $n/N \to p \in (0,1)$, avec $N = n+m$. Le paramètre p est interprété comme la proportion asymptotique des X dans l'échantillon regroupé. Nous définissons les statistiques de rang univariées basées sur l'échantillon regroupé, de distribution de mélange asymptotique égale à $F = pG + (1-p)H$. Sous des hypothèses classiques sur les distributions sous-jacentes, et en tenant compte d'éventuelles égalités, nous choisissons la définition des *upranks* (voir van der Vaart (1998), page 173), comme suit

$$\text{Rank}(t) = \sum_{i=1}^{n} \mathbb{I}\{X_i \leq t\} + \sum_{j=1}^{m} \mathbb{I}\{Y_j \leq t\}, \quad \text{for all} \quad t \in \mathscr{X} . \tag{C.2.3}$$

Les rangs étant basés sur l'échantillon regroupé, ils sont proportionnels à la distribution empirique de mélange F. Précisément, en considérant les versions empiriques des fonctions de répartitions G et H, *i.e.* , $\widehat{G}_n(t) = (1/n)\sum_{i \leq n} \mathbb{I}\{X_i \leq t\}$ et $\widehat{H}_m(t) = (1/m)\sum_{j \leq m} \mathbb{I}\{Y_j \leq t\}$, pour tout $t \in \mathscr{X}$, son estimateur est donné par $\widehat{F}_N(t) = (n/N)\widehat{G}_n(t) + (m/N)\widehat{H}_m(t)$, puisque $n/N \to p$ lorsque N tend vers l'infini. Nous obtenons alors

$$\text{Rank}(t) = N\widehat{F}_N(t), \quad \text{pour tout} \quad t \in \mathscr{X} .$$

Plus généralement, cette définition permet l'utilisation de quantiles empiriques, de copules, *etc.* Les R-statistiques linéaires à deux échantillons sont construites pouvant générer divers types de tests statistiques. Seuls les rangs des X parmi les échantillons regroupés sont considérés et adaptés/pondérés à l'aide d'une *fonction génératrice de score* $\phi : [0,1] \longrightarrow \mathbb{R}$, formellement définie comme l'une des représentations possibles des scores induits par les rangs, voir Def. (B.1.3) et (B.1.4) pour les principaux concepts générateurs.

Definition 98. (R-STATISTIQUES LINÉAIRES À DEUX ÉCHANTILLONS) *Soit $\phi : [0,1] \to \mathbb{R}$ une fonction strictement croissante. Les statistiques linéaires de rang à deux échantillons de "fonction géné-ratrice de score" $\phi(u)$ basée sur les échantillons $\{X_1, \ldots, X_n\}$ et $\{Y_1, \ldots, Y_m\}$ sont définies par :*

$$\widehat{W}_{n,m}^{\phi} = \sum_{i=1}^{n} \phi\left(\frac{\text{Rank}(X_i)}{N+1}\right). \tag{C.2.4}$$

Des résultats fondamentaux sur les R-statistiques à deux échantillons ont été obtenus grâce à H. Chernoff et I.R. Savage, et sur leur version généralisée par J. Hájek. Chernoff and Savage (1958) fournit une analyse asymptotique en écrivant les statistiques sous forme de mesures empiriques à l'aide de méthodes issues de von Mises. Hájek and Sidák (1967) formalise les propriétés essentielles et des exemples de R-statistiques. Dwass (1956) a proposé pour les R-statistiques linéaires à deux échantillons une formulation grâce aux statistiques non-biaisées (*unbiased*, U-). Il a prouvé leur

distribution limite gaussienne lorsque toutes les variables sont identiquement distribuées et, en particulier, a étudié la puissance asymptotique pour le test de localisation (Exemple 97). Ces résultats constituent les bases pour l'étude des statistiques linéaires de rang (Eq. (C.2.4)). Dans le contexte des tests à deux échantillons, leur propriété fondamentale est leur indépendance par rapport aux distributions sous-jacentes sous $\mathscr{H}_0$, voir Annexe B, Lemme 87-(ii). Le calcul exact des valeurs critiques sans aucune hypothèse de régularité sur les probabilités sous-jacentes des observations est alors possible. Nous nous référons à la section annexe B pour les propriétés fondamentales de ces statistiques univariées, à la fois sous l'hypothèse d'égalité des distributions et sous les alternatives. Ci-dessous, la Figure C.3 rassemble quelques choix pour ϕ conduisant à des tests classiques à deux échantillons, qui sont notamment détaillés dans la section Annexe B.

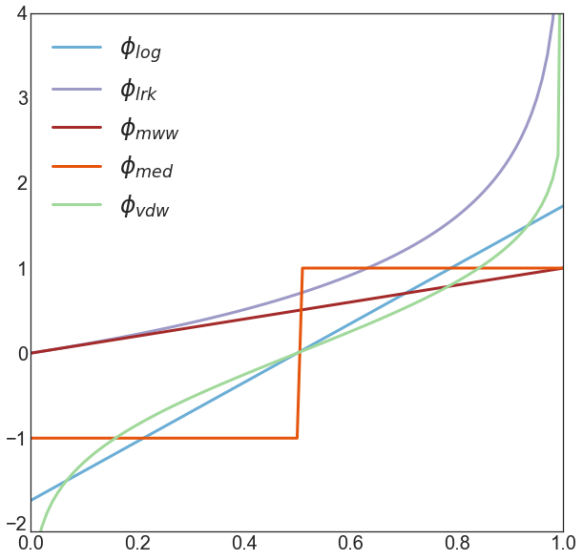


FIGURE C.3. Courbes de fonctions génératrices de score avec test de comparaison associé : test logistique $\phi_{log}(u) = 2\sqrt{3}(u-1/2)$ en bleu, test logrank $\phi_{lrk}(u) = -\log(1-x)$ en violet, test de Mann- Whitney-Wilcoxon $\phi_{mww}(u) = u$ en rouge, test de la médiane $\phi_{med}(u) = \mathrm{sgn}(u-1/2)$ en orange, test de Van der Waerden $\phi_{vdw}(u) = \Phi^{-1}(u)$ en vert, Φ est le quantile de la loi normale standard.

Méthodes d'apprentissage statistique d'ordonnancement

Les méthodes d'apprentissage automatique d'ordonnancement sont des approches statistiques de *classement* basées sur les données. Le but est d'apprendre une relation d'ordre à partir d'un ensemble d'observations selon leur pertinence/importance/préférence, pour prédire le rang de tout nouvel ensemble d'observations. La tâche d'apprentissage est formulée comme celle d'ordonnancement dans un cadre non/semi/-supervisé. Les applications sont nombreuses, telles que la recherche d'information, le Data Mining mais aussi les systèmes de recommandation (recherche web, listes préférentielles d'envoi, *etc.*) et les moteurs de recherche. Le contexte général des méthodes de classement est d'abord décrit. Ensuite, la formulation probabiliste des modèles appariés est détaillée, connue sous le nom de *classement bipartite*.

Dans l'approche la plus générique, le but est d'apprendre à classer un ensemble d'observations sachant un ensemble de requêtes, afin de minimiser un risque statistique. Cela revient à apprendre une fonction de scoring $s(z)$ définie sur un espace de caractéristiques multi-dimensionnel $\mathscr{Z}$ et à valeurs dans $\mathbb{R}$, telle que l'on peut classer toute paire d'observations : $a \preccurlyeq b$ *ssi* $s(a) < s(b)$ où $<$ est la relation d'ordre classique dans $\mathbb{R}$. Par exemple, dans la recherche de documents, le but est

d'apprendre un ordre parmi un ensemble de documents par rapport à leur pertinence, étant donné une série de requêtes. Ces problèmes d'optimisation ne considèrent les observations qu'à travers leur *rang*, au sens des statistiques d'ordre (Section C.2.2). La question fondamentale de ces modèles est d'apprendre à comparer des observations multivariées.

De nombreuses méthodes d'ordonnancement existent dépendamment des structures des données. D'une part, les méthodes d'apprentissage point à point résolvent le problème du classement en fonction de la pertinence de chaque élément. La fonction de perte évalue ensuite la qualité de la fonction de *scoring* apprise en comparant la prédiction de chaque donnée par rapport à l'ordre connu. Le classement peut donc être modélisé *via* la classification, la régression ou la régression ordinale. D'autre part, les méthodes appariées et plus largement par liste, formulent la fonction de perte basée sur la comparaison par paire/liste des éléments. Cette fonction mesure la précision des paires/listes d'observations prédites par une fonction de *scoring* par rapport à l'ordre connu. Les algorithmes associés sont plus complexes, étudiant au moins des comparaisons relatives par paires évaluées à toutes les données. Voir Liu (2009) pour une revue des modèles appliqués à la recherche d'information.

Ce manuscrit se concentre sur une approche binaire formulée comme un modèle d'apprentissage d'ordonnancement apparié avec requête unique, et connue sous le nom de modèle de *classement bipartite*. Les observations sont labélisées par une variable binaire, définie comme "positive" ou " négative ". Le but est d'apprendre leurs images univariées obtenues grâce à une fonction de *scoring* $s(z)$ optimale. Cette fonction induit alors un ordre minimisant la perte statistique de classement bipartite. Ces modèles se sont récemment développés à la détection d'anomalies, où le classement est appris afin d'ordonner les données en fonction de leur degré d'*anormalité*, voir Clémençon and Jakubowicz (2013); Clémençon and Thomas (2018); Frery et al. (2017); Müller et al. (2013). Voir la Section 2.3, Chap. 2, pour plus de détails sur ces dernières.

Formulation probabiliste des problèmes d'ordonnancement bipartites. Soit une variable $\mathbf{Z}$ d'entrée définiée sur l'espace de probabilité $(\Omega, \mathscr{A}, \mathbb{P})$ et à valeurs dans l'espace multivarié $\mathscr{Z}$, associée à son étiquette binaire ζ à valeurs dans $\{-1, +1\}$. L'ordonnancement bipartite peut être reformulé comme la comparaison de deux paires de variables aléatoires $(\mathbf{Z}, \zeta)$ et $(\mathbf{Z}', \zeta')$ conditionnellement à l'évènement $\{\zeta = 1, \zeta' = -1\}$ grâce à leur image par une fonction de *scoring*. La fonction optimale s^* est apprise parmi une classe de candidats $\mathscr{S} = \{s : \mathscr{Z} \to \mathbb{R} \cup \{+\infty\}, \ s \text{ mesurable}\}$, telle qu'elle minimise le risque d'ordonnancement bipartite, défini par :

$$L(s) = \mathbb{E}[\mathbb{I}\{s(\mathbf{Z}') > s(\mathbf{Z})\} \mid \zeta' = -1, \ \zeta = 1] + \frac{1}{2}\mathbb{P}\{s(\mathbf{Z}') = s(\mathbf{Z}) \mid \zeta' = -1, \ \zeta = 1\} , \qquad \text{(C.2.5)}$$

où les cas d'égalités sont tirés aléatoirement. De fait, s^* est définie par $L(s^*) = \inf_{\mathscr{S}} L =: L^*$. En considérant la probabilité a posteriori $\eta(z) = \mathbb{P}\{\zeta = 1 \mid \mathbf{Z} = z\}$, l'ensemble d'élèments optimaux est défini par, voir Clémençon and Vayatis (2008), Proposition 2 :

$$\mathscr{S}^* = \left\{ s \in \mathscr{S} \ \ s.t. \text{ pour tout } z, z' \text{ dans } \mathscr{Z} : \ \ \eta(z) < \eta(z') \Rightarrow s^*(z) < s^*(z') \right\} . \qquad \text{(C.2.6)}$$

Voir Clémençon and Vayatis (2008) pour les résultats d'optimalité associés. L'excès de risque pour une fonction $s(z)$ est défini par

$$L(s) - L^* = \mathbb{E}[|\eta(\mathbf{Z}') - \eta(\mathbf{Z})| \mathbb{I}\{(s(\mathbf{Z}) - s(\mathbf{Z}'))(\eta(\mathbf{Z}) - \eta(\mathbf{Z}')) < 0\}] , \qquad \text{(C.2.7)}$$

voir l'exemple 1 dans Clémençon et al. (2008). Cette formulation est fondamentale pour la compréhension de nombreuses méthodes de l'état de l'art, comme discuté en détails au Chap. 2. En pratique, la distribution sous-jacente étant inconnue, on considère la forme statistique basée sur des observations aléatoires *i.i.d.* $\{(\mathbf{Z}_i, \zeta_i)_{i \leq N}\}$, avec $N \in \mathbb{N}^*$. Le but de l'ordonnancement bipartite est donc

d'apprendre un *score* pour tout nouvel échantillon $\mathbf{Z}_{N+1}, \ldots, \mathbf{Z}_{N+k}$ de label inconnu, de sorte qu'il minimise la version empirique de la fonction de perte $L(s)$, définie par :

$$\widehat{L}(s) = \frac{1}{nm} \sum_{\{i,\, \zeta_i=+1\}} \sum_{\{j,\, \zeta_j=-1\}} \left(\mathbb{I}\{s(\mathbf{Z}_j) > s(\mathbf{Z}_i)\} + \frac{1}{2}\mathbb{I}\{s(\mathbf{Z}_j) = s(\mathbf{Z}_i)\} \right) , \tag{C.2.8}$$

où $n = \sum_{i \leq N} \mathbb{I}\{\zeta_i = +1\}$ et $m = \sum_{i \leq N} \mathbb{I}\{\zeta_i = -1\}$. La fonction de scoring optimale reproduit idéalement l'ordre induit par η et maximise les scores des observations 'positives' par rapport aux 'négatives'. Nous renvoyons à Menon and Williamson (2016) pour une revue détaillée des approches théoriques du classement bipartite et des algorithmes de l'état-de-l'art.

C.2.3 Organisation du manuscrit

Problèmes existants à deux échantillons et formulations mathématiques fondamentales. Cette partie regroupe et définit les concepts principaux traités dans cette thèse. Elle est divisée en deux chapitres :

- *Chapitre 2 : Problèmes à deux échantillons.* Tout d'abord, l'état de l'art des tests de comparaison multivariés et non paramétriques à deux échantillons est étudié. Ensuite, deux modèles d'apprentissage d'ordonnancement sont détaillés, à savoir les problèmes d'ordonnancement bipartite et d'anomalie. Tous sont formulés sous leur forme générique, tout en passant en revue les principaux résultats et les méthodes de référence.

- *Chapitre 3 : Quelques résultats de concentration.* Motivés par la théorie de minimisation du risque empirique, la construction d'inégalités de concentration est présentée. Cela permet le contrôle de statistiques et surtout de leurs collections, définies comme processus empiriques (si statistiques d'ordre 1), U-processus (si d'ordre supérieur).

Contributions relatives aux R-processus. Cette deuxième partie constitue le cœur de la thèse, développant une analyse de la version généralisée des R-processus et de leur application au problème à deux échantillons. Les simulations numériques sur des données synthétiques sont rassemblées dans le dernier chapitre.

- *Chapitre 4 : Une inégalité de concentration pour U-processus.* Ce chapitre d'introduction démontre une nouvelle concentration pour des U-processus dégénérés particuliers à deux échantillons, lorsqu'il sont indexés par une classe de noyaux de complexité contrôlée. C'est un nouveau résultat pour la littérature, et nécessaire pour le chapitre suivant. Il a été publié dans le cadre de l'article [1].

- *Chapitre 5 : Inégalités de concentration pour des R-processus basés sur deux échantillons.* Les résultats relatifs aux R-processus sont démontrés et motivés par les modèles d'ordonnancement bipartite. Ce chapitre correspond à la publication [1].

- *Chapitre 6 : Tests statistiques d'homogénéité à deux échantillons.* Une formulation générique du problème à deux échantillons basée sur les R-statistiques est proposée, optimisée grâce à des algorithmes d'apprentissage d'ordonnancement. Nous énonçons une procédure en deux étapes avec garanties théoriques. Il s'agit d'un document en cours de travail.

- *Chapitre 7 : Simulations numériques.* Cette section rassemble des expériences numériques basées sur des données synthétiques afin de tester nos critères basés sur le rang proposés dans deux contextes : l'ordonnancement bipartite et le test de comparaison à deux échantillons.

Tous les détails sur les codes sont en particulier détaillés. Ce chapitre regroupe les résultats numériques de [1] (Chapitre 6) et [2] (Chapitre 8). Les algorithmes sont codés en `Python` et sont en lignès en accès libre à `https://github.com/MyrtoLimnios`.

Applications. Cette dernière partie se concentre sur trois contributions applicatives. Alors que la première est liée à un modèle d'apprentissage du rang, les deux suivantes sont issues de la recherche interdisciplinaire effectuée au Centre Borelli et liées à l'analyse et modélisation du contrôle postural.

- *Chapitre 8 : Apprentissage d'ordonnancement d'anomalies avec des R-statistiques basées sur deux échantillons.* Nous proposons une méthodologie pour apprendre à classer les observations selon leur *degré* d'anormalité. La publication associée est [3].

- *Chapitre 9 : Tests de comparaisons appliqués aux études biomédicales.* Une méthode de test d'homogénéité à deux échantillons pour les applications biomédicales est détaillée, adaptée pour maximiser une version particulière des R-statistiques proposées. Cet algorithme est appliqué à la comparaison statistique de deux populations cliniques, et plus particulièrement basé sur des mesures extraites de statokinésigrammes. Les publications/communications associées sont [4-5].

- *Chapitre 10 : Un modèle génératif pour le contrôle posutral.* Un modèle est proposé pour générer l'évolution temporelle du centre de pression lorsqu'il est modélisé par une correlation temporelle au centre de masse. Ce modèle est basé sur le modèle stochastique de Langevin. La publication associée est [6].

Appendice. L'appendice regroupe trois chapitres comme suit.

- *Appendice A : R-processus à deux échantillons généralisés et tests de comparaison efficients.* Cette section est liée à l'étude des R-statistiques lorsqu'elles sont indexées par une classe de fonctions génératrices de score ϕ. Dans la continuité du chapitre 5 et inspirés des travaux de H. Koul (voir Section 1.3.2), nous étudions les R-processus sous des hypothèses faibles supposées pour la fonction génératrice de score. Ensuite, pour les tests de comparaison, une procédure supplémentaire est décrite, dans laquelle *Étape 1.* de la Fig. 1.4 est remplacée par la maximisation exacte de la R-statistique. Nous utilisons l'algorithme 1 et quelques expériences numériques sont fournies. En outre, une approche adaptative pour choisir la "meilleure" fonction génératrice de score est détaillée.

- *Appendice B : Formulation univariée et état de l'art.* Cette section développe des résultats fondamentaux et des exemples sur les problèmes univariés suivants : les statistiques de rang à deux échantillons, les tests de comparaison à deux échantillons et l'analyse ROC.

- *Appendice C : Suppléments.* Certains faits sur la recherche scientifique basée sur des statistiques sont traités et en particulier concernant la recherche reproductible. Enfin, l'introduction générale du manuscrit en français est dévelopée.

université PARIS-SACLAY | ÉCOLE DOCTORALE de mathématiques Hadamard (EDMH)

Titre: Processus de Rang et Applications Statistiques en Grande Dimension
Mots clés: statistiques linéaires de rang, apprentissage statistique, inégalités de concentration, processus de rang, tests de comparaison, problème à deux échantillons

Résumé: Ce projet de recherche propose de développer des outils mathématiques et algorithmiques pour étudier et comparer deux jeux de données complexes en grande dimension : vecteurs, signaux multivariés, trajectoires, signaux sur graphes. Il répond à des enjeux fondamentaux liés à la quantification dans les sciences expérimentales, notamment les sciences de la vie et par-là même les neurosciences et ses applications cliniques. Pour se faire, nous proposons une généralisation des statistiques linéaires de rang à l'aide d'outils développés en apprentissage automatique. En effet, et grâce à des techniques d'ordonnancement bipartite, une étude avancée et non-paramétrique de ces statistiques à deux échantillons est menée sous l'angle de la théorie de l'apprentissage statistique. Plus précisément, ces méthodes permettent de pallier l'absence de relation d'ordre dans les espaces de grande dimension grâce à l'apprentissage d'une fonction de score. Définie sur l'espace ambiant et à valeur réelle, cette dernière a pour but d'induire un ordre sur les observations multivariées en maximisant la statistique de rang généralisée. Nous proposons une première application dans le cadre des tests d'hypothèses statistiques, en associant décision (acceptation/rejet) de l'hypothèse nulle à l'apprentissage d'un modèle décrivant les données. Nous étudions, plus précisément, les tests d'homogénéité à deux échantillons. Ensuite, deux applications en analyse de données sont introduites et développées en utilisant les statistiques de rang comme critère scalaire de performance. Nous les appliquons aux problèmes d'ordonnancement bipartite et d'apprentissage des données extrêmes, ou anomalies, et précisons leurs relations à l'état de l'art. Enfin, dans la volonté de proposer des outils adaptés aux données issues des sciences expérimentales et dans le cadre de l'étude des données biomédicales, nous introduisons une méthode interprétable de comparaison statistique de deux populations cliniques, ainsi que d'un modèle stochastique génératif de données longitudinales particulières.

Title: Rank Processes and Statistical Applications in High Dimension
Keywords: linear rank statistics, statistical learning, concentration bounds, rank processes, homogeneity testing, two-sample problem

Abstract: This research project aims at developing mathematical and algorithmic tools to study and evaluate the level of similarity between two complex datasets in high-dimension: vectors, multivariate signals, trajectories, signals on graphs. It answers fundamental questions related to quantification in experimental science, particularly in life sciences, neurosciences, and clinical applications. We propose a generalization of linear rank statistics using methods developed in machine learning. Indeed, thanks to bipartite ranking approaches, we articulate an in-depth and nonparametric study of those statistics based on two statistical samples, using statistical learning theory. More precisely, ranking methods circumvent the lack of relation order in high-dimensional spaces by learning a scoring function. The latter, defined on the ambient space and valued in the real line, aims at inducing an order on the multivariate observations by maximizing the generalized rank statistic. We propose the first application in statistical hypothesis testing by combining decision (acceptance/rejection) of the null hypothesis and learning a model describing the data. More specifically, we study two-sample homogeneity tests. Then, two applications in data analysis are introduced and developed using rank statistics as a performance criterion. They are applied to bipartite ranking and anomaly detection problems and specify their relation to state-of-the-art formulations. Finally, and motivated to propose tools adapted to experimental sciences and in the context of biomedical data studies, we introduce an interpretable method for the statistical comparison of two clinical populations and a stochastic generative model of specific longitudinal data.


école
normale
supérieure
paris–saclay


université
PARIS-SACLAY


CENTRE
BORELLI


Fondation mathématique
FMJH
Jacques Hadamard


MATH
INNOV


îledeFrance