Statistical Depth Functions for Ranking Distributions: Definitions, Statistical Learning and Applications

Morgane Goibert
Télécom Paris
Criteo

Stéphan Clémençon
Télécom Paris

Ekhine Irurozki
Télécom Paris

Pavlo Mozharovskyi
Télécom Paris

Abstract

The concept of *median/consensus* has been widely investigated in order to provide a statistical summary of ranking data, *i.e.* realizations of a random permutation Σ of a finite set, $\{1, \ldots, n\}$ with $n \geq 1$ say. As it sheds light onto only one aspect of Σ's distribution P, it may neglect other informative features. It is the purpose of this paper to define analogues of quantiles, ranks and statistical procedures based on such quantities for the analysis of ranking data by means of a metric-based notion of *depth function* on the symmetric group. Overcoming the absence of vector space structure on $\mathfrak{S}_n$, the latter defines a center-outward ordering of the permutations in the support of P and extends the classic metric-based formulation of *consensus ranking* (*medians* corresponding then to the *deepest* permutations). The axiomatic properties that *ranking depths* should ideally possess are listed, while computational and generalization issues are studied at length. Beyond the theoretical analysis carried out, the relevance of the novel concepts and methods introduced for a wide variety of statistical tasks are also supported by numerous numerical experiments.

1 Introduction

The statistical analysis of ranking data as recently received much attention (*e.g.* Alvo and Yu (2014) and references therein), fed by the increasing number of modern applications involving *preferences* data (search engines, recommender systems, etc.). Such data usually consist of $N \geq 1$ permutations $\sigma_1, \ldots, \sigma_N$ on an ensemble of $n \geq 1$ items, indexed by $i \in \{1, \ldots, n\}$. The major scientific challenge arises from *the absence of any vector space structure* on the set of all permutations, the symmetric group $\mathfrak{S}_n$. Given the impossibility of 'averaging' the σ_j's in a straightforward manner, the issue of summarizing a ranking dataset by a single permutation, referred to as *Consensus Ranking* or *Ranking Aggregation*, has concentrated much interest (seminal works of de Borda (1781); Condorcet (1785) in social choice theory, Patel et al. (2013) in bioinformatics, Desarkar et al. (2016) in meta-search engines, Davenport and Lovell (2005) in competition ranking, etc.). Two approaches to Consensus Ranking have been studied. The first one, initiated by Condorcet in the 18th century, is based on probabilistic modelling. The second one is a metric-based: equipped with a (pseudo-) distance on $\mathfrak{S}_n$, a barycentric permutation, referred to as a *ranking median*, is found. However, central measures such as medians shed light on only one aspect of a multivariate distribution and ignore other interesting characteristics. Thus, the informative nature of ranking medians about the distribution P of a random permutation Σ, *i.e.* a r.v. takings its values in $\mathfrak{S}_n$, is limited and must be complemented by additional quantities, providing information analogous to that illuminated by quantiles for a univariate distribution.

This article is devoted to defining such quantities for ranking data. We extend the statistical depth concept, originally introduced so as to define quantiles for probability distributions on $\mathbb{R}^d$ with $d \geq 2$ (see *e.g.* Mosler (2013)), to ranking distributions. Some basics in statistical depth theory are briefly recalled in section 2, while section 3 introduced an extension of the notion of depth function tailored to ranking data. Desirable properties for ranking depths are listed therein, and shown to hold under mild conditions, *e.g.* stochastic transitivity. Based on a pseudo-metric on $\mathfrak{S}_n$, the depth of a ranking σ relative to P measures its expected closeness to the

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

random permutation Σ. Hence, ranking medians correspond to the deepest rankings. In section 4, statistical guarantees are provided for the ranking depth and its by-products, in the form of non-asymptotic bounds for the deviations between the ranking depth function and its statistical counterpart in particular. A trimming algorithm, based on the ranking depth concept, to recover automatically a stochastically transitive version of the empirical ranking distribution is also proposed therein. Beyond the theoretical/algorithmic concepts introduced and analyzed here, the relevance of the notion of ranking depth is motivated by a wide variety of statistical applications, illustrated by several numerical experiments in section 5.

The main contributions of the paper are summarized below:

- Statistical depth and related axiomatic properties are extended to ranking data, in order to emulate quantiles/ranks for r.v.'s valued in $\mathfrak{S}_n$.
- A finite-sample analysis ensures the usability of the notion of ranking depth introduced.
- An algorithm of great simplicity that uses ranking depth to build stochastically transitive empirical ranking distributions (based on which, crucial statistical tasks such as consensus ranking are straightforward) is proposed.
- The ranking depth and the related quantile regions in $\mathfrak{S}_n$ it defines can be used for the statistical analysis of rankings: 1) fast and robust recovery of medians in consensus ranking, 2) informative graphical representations of ranking data, 3) anomaly/novelty detection, 4) homogeneity testing.

2 Background and Preliminaries

We start with recalling some basics in statistical depth theory, together with key notions of the statistical analysis of ranking data involved in the subsequent analysis. Throughout the paper the indicator function of any event $\mathcal{E}$ is denoted by $\mathbb{I}\{\mathcal{E}\}$, the Dirac mass at any point a by δ_a, the floor function by $u \in \mathbb{R} \mapsto \lfloor u \rfloor$, the convolution product of two real valued functions f and g defined on the real line, when well-defined, by $f * g$, the cardinality of any finite set E by $\#E$ and the set of permutations of $\{1, \ldots, n\}$ by $\mathfrak{S}_n$ for $n \geq 1$.

2.1 Depth Functions for Multivariate Data

In absence of any 'natural order' on $\mathbb{R}^d$ with $d \geq 2$, the concept of *statistical depth* permits to define a center-outward ordering of points in the support of a probability distribution P on $\mathbb{R}^d$, so as to extend the notions of order and (signed) rank statistics to multivariate data, see *e.g.* Mosler (2013). A depth function $D_P : \mathbb{R}^d \to \mathbb{R}_+$ relative to P should ideally assign the highest values $D_P(x)$ to points $x \in \mathbb{R}^d$ near the "center" of the distribution. Originally introduced in the seminal contribution Tukey (1975), the *half-space depth* of x in $\mathbb{R}^d$ relative to P is the minimum of the mass $P(H)$ taken over all closed half-spaces $H \subset \mathbb{R}^d$ such that $x \in H$. Many alternatives have been proposed since then, see *e.g.* Liu (1990); Liu and Singh (1993); Koshevoy and Mosler (1997); Chaudhuri (1996); Oja (1983); Vardi and Zhang (2000); Chernozhukov et al. (2017); Zuo and Serfling (2000a). To compare the merits and drawbacks of different notions of depth function, an axiomatic nomenclature has been introduced in Zuo and Serfling (2000a), listing four properties that statistical depths should ideally satisfied, see Dyckerhoff (2004); Mosler (2013) for a different formulation of a statistically equivalent set of properties.

(*i*) (Affine invariance) Denoting by P_X the distribution of any r.v. X taking its values in $\mathbb{R}^d$, it holds: $D_{P_{AX+b}}(Ax + b) = D_P(x)$ for all $x \in \mathbb{R}^d$, any r.v. X valued in $\mathbb{R}^d$, any $d \times d$ nonsingular matrix A with real entries and any vector b in $\mathbb{R}^d$.

(*ii*) (Maximality at center) For any probability distribution P on $\mathbb{R}^d$ that possesses a symmetry center x_P (for different notions of center), the depth function D_P takes its maximum value at it, *i.e.* $D_P(x_P) = \sup_{x \in \mathbb{R}^d} D_P(x)$.

(*iii*) (Monotonicity relative to deepest point) For any probability distribution P on $\mathbb{R}^d$ with deepest point x_P, the depth at any point x in $\mathbb{R}^d$ decreases as one moves away from x_P along any ray passing through it, *i.e.* $D_P(x) \leq D_P(x_P + \alpha(x - x_P))$ for any α in $[0, 1]$.

(*iv*) (Vanishing at infinity) For any probability distribution P on $\mathbb{R}^d$, the depth function D_P vanishes at infinity, *i.e.* $D_P(x) \to 0$ as $||x||$ tends to infinity.

As the distribution P of interest is generally unknown in practice, its analysis relies on the observation of $N \geq 1$ independent realizations $X_1, \ldots, X_N$ of P. A statistical version of $D_P(x)$ can be built by replacing P with its empirical counterpart $\widehat{P}_N = (1/N) \sum_{i=1}^N \delta_{X_i}$, yielding the *empirical depth function* $D_{\widehat{P}_N}(x)$. Its consistency and asymptotic normality have been studied for various notions of depth, refer to *e.g.* Donoho and Gasko (1992); Zuo and Serfling (2000b), and concentration results for empirical depth and contours have been recently proved in the half-space depth case, see Burr and Fabrizio (2017); Brunel (2019).

2.2 Consensus Ranking

Given a certain metric $d(.,\ .)$ on $\mathfrak{S}_n$ and a r.v. Σ defined on a probability space $(\Omega,\ \mathcal{F},\ \mathbb{P})$ and drawn from an unknown probability distribution P on $\mathfrak{S}_n$ (*i.e.* $P(\sigma) = \mathbb{P}\{\Sigma = \sigma\}$ for any $\sigma \in \mathfrak{S}_n$), the metric approach to consensus ranking consists in finding a ranking $\sigma^* \in \mathfrak{S}_n$ whose expected distance to Σ is minimum, *i.e.* such that

$$L_P(\sigma^*) = \min_{\sigma \in \mathfrak{S}_n} L_P(\sigma), \tag{1}$$

where $L_P(\sigma) = \mathbb{E}_P[d(\Sigma, \sigma)]$ is referred to as the *ranking risk* of any median candidate σ in $\mathfrak{S}_n$ w.r.t. d and Σ. The ranking median σ^* (not necessarily unique) is viewed as an informative summary of P and $L_P(\sigma^*)$ as a dispersion measure. The choice of the (pseudo) distance $d(.,\ .)$ is crucial, regarding the theoretical properties of the corresponding medians and the computational feasibility, see section 3. Various distances have been considered in the literature, see *e.g.* Deza and Huang (1998), the most popular choices being listed below: $\forall (\sigma, \sigma') \in \mathfrak{S}_n^2$,

$$\begin{aligned}
d_\tau(\sigma, \sigma') &= \sum_{i<j} \mathbb{I}\left\{(\sigma(i) - \sigma(j))(\sigma'(i) - \sigma'(j)) < 0\right\}, \\
d_2(\sigma, \sigma') &= \left(\sum_{i=1}^n (\sigma(i) - \sigma'(i))^2\right)^{1/2}, \\
d_1(\sigma, \sigma') &= \sum_{i=1}^n |\sigma(i) - \sigma'(i)|, \\
d_H(\sigma, \sigma') &= \sum_{i=1}^n \mathbb{I}\left\{\sigma(i) \neq \sigma'(i)\right\},
\end{aligned}$$

known respectively as the Kendall τ, the Spearman ρ, the Spearman footrule and the Hamming distances. The literature has essentially focused on solving a statistical version of the minimization problem (1), see *e.g.* Hudry (2008), Diaconis and Graham (1977) or Bartholdi III et al. (1989). Assuming that $N \geq 1$ independent copies $\Sigma_1,\ \ldots,\ \Sigma_N$ of the generic r.v. Σ are observed, a natural empirical estimate of $L_P(\sigma)$ is $\widehat{L}_N(\sigma) = (1/N)\sum_{s=1}^N d(\Sigma_s, \sigma) = L_{\widehat{P}_N}(\sigma)$, where $\widehat{P}_N = (1/N)\sum_{i=1}^N \delta_{\Sigma_i}$ is the empirical measure. The set $\mathfrak{S}_n$ being of finite cardinality, an empirical ranking risk minimizer always exists, just like a solution to (1), not necessarily unique however. Generalization guarantees and fast rate conditions for empirical consensus ranking have been investigated in Clémençon et al. (2017).

3 Depth Functions for Ranking Data

In order to define relevant extensions of the concept of statistical depth to ranking data, we define axiomatic properties that candidate functions on $\mathfrak{S}_n$ should satisfy. We next show that the metric-based ranking depths we propose to analyze ranking distributions satisfy these properties under mild conditions.

3.1 Ranking Depths - Axioms

Just like in the multivariate setup (see subsection 2.1), a list of key properties the ranking depth function D_P should ideally satisfy can be made. These properties are essential to emulate the information provided by quantiles (resp. quantile regions) of univariate distributions (resp. multivariate distributions) in a relevant manner. Let P be a ranking distribution, d a distance on $\mathfrak{S}_n$, the properties desirable for any ranking depth $D_P : \mathfrak{S}_n \to \mathbb{R}_+$ are listed below.

Property 1. (Invariance) *For any $\pi \in \mathfrak{S}_n$, consider the ranking distribution πP defined by: $(\pi P)(\sigma) = P(\sigma\pi^{-1})$ for all $\sigma \in \mathfrak{S}_n$. It holds that: $D_P(\sigma) = D_{\pi P}(\sigma\pi)$ for all $(\sigma,\ \pi) \in \mathfrak{S}_n^2$.*

Property 2. (Maximality at center) *For any probability distribution P on $\mathfrak{S}_n$ that possesses a symmetry center σ_P (in a certain sense, e.g. w.r.t. to a given metric d on $\mathfrak{S}_n$), the depth function D_P takes its maximum value at it, i.e. $D_P(\sigma_P) = \max_{\sigma \in \mathfrak{S}_n} D_P(\sigma)$.*

Property 3. (Local monotonicity relative to deepest ranking) *Assume that the deepest ranking σ^* is unique. The quantity $D_P(\sigma)$ decreases as $d(\sigma^*, \sigma)$ locally increases, i.e. for any π such that $d(\sigma^*, \sigma\pi) = d(\sigma^*, \sigma) + 1$, then we have $D_P(\sigma) > D_P(\sigma\pi)$.*

Note that, insofar as $\mathfrak{S}_n$ is of finite cardinality, there is no relevant analogue of the 'vanishing at infinity' property for multivariate depth. A stronger monotonicity property can also be formulated.

Property 4. (Global monotonicity) *Assume that the deepest ranking σ^* is unique. The quantity $D_P(\sigma)$ decreases as $d(\sigma^*, \sigma)$ globally increases, i.e. $d(\sigma^*, \sigma') > d(\sigma^*, \sigma) \Rightarrow D_P(\sigma') < D_P(\sigma)$.*

3.2 Metric-based Ranking Depth Functions

Seeking to define a ranking depth that satisfies the properties listed above and such that the medians σ_P^* of P have maximal depth, the metric approach provides natural candidates, just like for consensus ranking.

Definition 1. (Metric-based ranking depth) Let d be a distance and P a distribution on $\mathfrak{S}_n$. The ranking depth based on d is defined as: $D_P^{(d)}$: $\forall \sigma \in \mathfrak{S}_n$, $D_P^{(d)}(\sigma) = \mathbb{E}_P[||d||_\infty - d(\sigma, \Sigma)] = ||d||_\infty - L_P(\sigma)$, with $||d||_\infty = \max_{(\sigma,\sigma') \in \mathfrak{S}_n^2} d(\sigma, \sigma')$.

The shift induced by $||d||_\infty \geq L^\star = \max_{\sigma\in\mathfrak{S}_n} L_P(\sigma)$ simply guarantees non-negativity, in accordance with Definition 2.1 in Zuo and Serfling (2000a), while defining the same center-outward ordering of the permutations σ in $\mathfrak{S}_n$ as $-L_P$. Notice that metric-based ranking depths can be viewed as extensions of multivariate depth functions of type A in the nomenclature proposed in Zuo and Serfling (2000a). For simplicity, we omit the superscript (d) and rather write D_P when no confusion is possible about the distance considered.

A ranking σ in $\mathfrak{S}_n$ is said to be *deeper* than another one σ' relative to the ranking distribution P iff $D_P(\sigma') \leq D_P(\sigma)$ and we write $\sigma' \preceq_{D_P} \sigma$. The *ranking depth ordering* $\preceq_{D_P}$ is the preorder related to the depth function D_P. Equipped with this notion of depth on $\mathfrak{S}_N$, medians σ^* of P w.r.t. the metric d correspond to the deepest rankings. If P is a Dirac mass δ_{σ_0}, the ranking depth then simply reduces to the measure of closeness defined by the distance d chosen: $D_P(\sigma) = ||d||_\infty - d(\sigma_0, \sigma)$. In contrast, if P is the uniform distribution, the ranking depth relative to a classic distance on $\mathfrak{S}_n$ is constant over $\mathfrak{S}_n$. The depth function also permits to partition the space $\mathfrak{S}_n$ into subsets of rankings with equal depth.

Definition 2. (Depth regions/contours) For any $u \in \mathbb{R}$, the region of depth u is the superlevel set $\mathcal{R}_P(u) = \{\sigma \in \mathfrak{S}_n : \ D_P(\sigma) \geq u\}$ of D_P, while the ranking contour of depth u is the set $\partial\mathcal{R}_P(u) = \{\sigma \in \mathfrak{S}_n : \ D_P(\sigma) = u\}$.

Equipped with this notation, $\partial\mathcal{R}_P(-L_P^*)$ is the set of medians of P w.r.t. the metric d.

Definition 3. (Depth survivor function) The ranking depth survivor function is $S_P : u \in \mathbb{R} \mapsto S_P(u) = \mathbb{P}\{D_P(\Sigma) \geq u\}$.

Based on the metric-based ranking depth, the quantile regions are defined as follows.

Definition 4. (Quantile regions in $\mathfrak{S}_n$) Let $\alpha \in (0,1)$. The depth region with probability content α is the region of depth $S_P^{-1}(\alpha) = \inf\{u \in \mathbb{R} : \ S_P(u) \leq 1-\alpha\}$: $R_P(\alpha) = \mathcal{R}_P(S_P^{-1}(\alpha))$. The mapping $\alpha \in (0,1) \mapsto S_P^{-1}(\alpha)$ is called the ranking quantile function.

3.3 The Metric Approach - Main Properties

We now state results showing that, under mild conditions and for popular choices of d, the metric-based ranking depth introduced in Definition 1 satisfies the key properties listed in subsection 3.1. Technical proofs are postponed to the Supplementary Material.

Proposition 1. (Invariance) *Suppose that d is right-invariant, i.e. $d(\nu\pi, \sigma\pi) = d(\nu, \sigma)$ for all $(\nu, \pi, \sigma) \in \mathfrak{S}_n^3$, the ranking depth $D_P^{(d)}$ satisfies the Property 1.*

We point out that Spearman ρ, Spearman footrule, Kendall τ, Hamming, Ulam and Cayley distances are all right-invariant. Hence, the invariance property is satisfied for any ranking distribution in many situations. Checking the other properties is more challenging. We recall the following notion.

Definition 5. (Stochastic transitivity) *A probability distribution P on $\mathfrak{S}_n$ is said to be stochastically transitive (ST) iff, for all $(i,j,k) \in [\![n]\!]^3$, we have: $p_{i,j} \geq 1/2$ and $p_{j,k} \geq 1/2 \ \Rightarrow \ p_{i,k} \geq 1/2$. If, in addition, $p_{i,j} \neq 1/2$ for all $i < j$, one says that P is strictly stochastically transitive (SST).*

The *stochastic transitivity* property (Fishburn, 1973; Davidson and Marschak, 1959) is fulfilled by some widely used ranking distributions (*e.g.* Mallows) and shown to facilitate various statistical tasks, see *e.g.* Shah et al. (2015); Shah and Wainwright (2015). In particular, if P is SST, Kemeny's median (*i.e.* the median σ^* w.r.t. Kendall τ distance) is unique, see *e.g.* Clémençon et al. (2017).

Proposition 2. (Maximality at the center)*: The Spearman's footrule ranking depth satisfies Property 2 for any distribution P with a symmetry center. If P is SST in addition, then Kendall τ and Spearman ρ ranking depths satisfy Property 2 as well.*

Proposition 3. (Local monotonicity) *If the distribution P is SST, then the Kendall τ ranking depth satisfies Property 3.*

Proposition 4. (Global monotonicity) *If the distribution P is SST and $||d_\tau||_\infty = \binom{n}{2} < h/s$ with $h = \min_{i,j} |p_{i,j} - 1/2|$ and $s = \max_{(i,j)\neq(k,l)} |p_{i,j} - p_{k,l}|$, then the Kendall τ ranking depth satisfies Property 4.*

In the Kendall τ case, additional useful results can be stated. In particular, the ranking depth is then entirely determined by the *pairwise probabilities* $p_{i,j} = \mathbb{P}\{\Sigma(i) < \Sigma(j)\}$, $1 \leq i \neq j \leq n$.

Proposition 5. *We have: $\forall\, \sigma \in \mathfrak{S}_n$, $D_P(\sigma) = \binom{n}{2} - \sum_{i<j} p_{i,j}\mathbb{I}\{\sigma(i) > \sigma(j)\} - \sum_{i<j}(1-p_{i,j})\mathbb{I}\{\sigma(i) < \sigma(j)\}$.*

This case is computationally attractive, the complexity being of order $O(n^2)$. In addition, note that the computation of D_P involves pairwise comparisons solely, which means an alternative statistical framework can be considered, where observation take the form of binary variables $\{\Sigma(\mathbf{i}) < \Sigma(\mathbf{j})\}$, $(\mathbf{i}, \mathbf{j})$ being a random pair in $\{(i,j) : \ 1 \leq i < j \leq n\}$, independent from Σ.

Proposition 6. *Suppose that the ranking distribution P is stochastichally transitive. The following assertions hold true.*

(i) The largest ranking depth value is $D_P^* = \sum_{i<j} \left\{\frac{1}{2} + \left|p_{i,j} - \frac{1}{2}\right|\right\}$. The deepest rankings relative to P and d_τ are the permutations $\sigma \in \mathfrak{S}_n$

such that: $\forall i < j$ *s.t.* $p_{i,j} \neq 1/2$, $(\sigma(j) - \sigma(i)) \cdot (p_{i,j} - 1/2) > 0$.

(ii) The smallest ranking depth value is $\underline{D}_P = \sum_{i<j} \left\{ \frac{1}{2} - \left| p_{i,j} - \frac{1}{2} \right| \right\}$. *The least deep rankings relative to* P *and* d_τ *are the permutations* $\sigma \in \mathfrak{S}_n$ *such that:* $\forall i < j$ *s.t.* $p_{i,j} \neq 1/2$, $(\sigma(j) - \sigma(i)) \cdot (p_{i,j} - 1/2) < 0$.

(iii) If, in addition, P *is SST, then we have* $\partial \mathcal{R}_P(D_P^*) = \{\sigma^*\}$ *and* $\partial \mathcal{R}_P(\underline{D}_P) = \{\underline{\sigma}\}$, *where* $\sigma^*(i) = 1 + \sum_{j \neq i} \mathbb{I}\{p_{i,j} < 1/2\} = n - \underline{\sigma}(i)$ *for* $i \in \{1, \ldots, n\}$. *We also have* $D_P^* - D_P(\sigma) = 2\sum_{i<j} |p_{i,j} - 1/2| + D_P(\sigma) - \underline{D}_P = 2\sum_{i<j} |p_{i,j} - 1/2| \cdot \mathbb{I}\{(\sigma(j) - \sigma(i))(p_{i,j} - 1/2) < 0\}$.

4 Statistical Issues

The ranking depth D_P is generally unknown, just like the ranking distribution P, and must be replaced by an empirical estimate based on supposedly available ranking data in practice. Here we establish nonasymptotic statistical guarantees for the empirical counterpart of the ranking depth and other related quantities. We also propose an algorithm, based on the ranking depth, that permits to build, from any ranking dataset, an empirical ranking distribution fulfilling the crucial (strict) stochastic transitivity property, see subsection 3.3.

4.1 Generalization - Learning Rate Bounds

Based on the observation of an i.i.d. sample $\Sigma_1, \ldots, \Sigma_N$ drawn from P with $N \geq 1$, statistical versions of the quantities introduced in subsection 3.2 can be built by replacing P with the empirical distribution $\widehat{P}_N$. The empirical ranking depth is thus given by: $\forall \sigma \in \mathfrak{S}_n$, $\widehat{D}_N(\sigma) = D_{\widehat{P}_N}(\sigma) = ||d||_\infty - \widehat{L}_N(\sigma)$. Similarly, the empirical ranking depth regions are $\widehat{\mathcal{R}}_N(u) = \{\sigma \in \mathfrak{S}_n : \widehat{D}_N(\sigma) \geq u\}$ for $u \geq 0$. In order to build an estimator of the ranking depth survivor function $S_P(u)$ with a tractable dependence structure, a *2-split* trick can be used, yielding the statistic

$$\widehat{S}_N(u) = \frac{1}{N - \lfloor N/2 \rfloor} \sum_{i=1+\lfloor N/2 \rfloor}^{N} \mathbb{I}\{\widehat{D}_{\lfloor N/2 \rfloor}(\Sigma_i) \geq u\}.$$

As the r.v. $D_P(\Sigma)$ is discrete, the use of smoothing/interpolation procedures is required to ensure good statistical properties for the survivor function estimator and for the empirical quantiles it defines, see Sheather and Marron (1990); Ma et al. (2011). For instance, a kernel smoothed version of S_P can be computed by means of a non-negative differentiable Parzen-Rosenblatt kernel $K : \mathbb{R} \to \mathbb{R}_+$ s.t. $||K'||_\infty = \sup_{u \in \mathbb{R}} |K'(u)| < \infty$ and $\int_{\mathbb{R}} K(u)du = +1$ and a smoothing bandwidth $h > 0$, namely: $\widetilde{S}_P(u) = K_h * S_P$, which can be estimated by $\widetilde{S}_N(u) = K_h * \widehat{S}_N$, where $K_h(u) = K(u/h)/h$ for $u \in \mathbb{R}$. One may then define a smooth estimate of the ranking depth region with probability content $\alpha \in [0, 1]$ as well: $\widehat{R}_N(\alpha) = \widehat{\mathcal{R}}_N(\widetilde{S}_N^{-1}(\alpha))$. The result below provides bounds of order $O_{\mathbb{P}}(1/\sqrt{N})$ for the maximal deviations between D_P (resp. $\widetilde{S}_P$) and its empirical version.

Proposition 7. *The following assertions hold true.*

(i) For any $\delta \in (0, 1)$, *we have with probability at least* $1 - \delta$: $\forall N \geq 1$,

$$\sup_{\sigma \in \mathfrak{S}_n} |\widehat{D}_N(\sigma) - D_P(\sigma)| \leq ||d||_\infty \sqrt{\frac{\log(2\, n!/\delta)}{2N}}.$$

(ii) For any $\delta \in (0, 1)$ *and* $h > 0$, *we have with probability at least* $1 - \delta$: $\forall N \geq 1$,

$$\sup_{u \geq 0} |\widetilde{S}_N(u) - \widetilde{S}_P(u)| \leq \sqrt{\frac{\log(4/\delta)}{2N}} + ||d||_\infty \sqrt{\frac{\log(4n!/\delta)}{2N}}.$$

For the technical proof, refer to the Supplementary Material, where the asymptotic rate for the empirical ranking quantile function is also given.

4.2 Depth Trimming for Consensus Ranking

As discussed in subsection 3.3, (strict) stochastic transitivity greatly facilitates the computation of Kemeny medians (see Proposition 6) as well as the verification of the maximality or monotonicity properties, *cf* Propositions 2, 3 and 4. However, although this occurs with a controlled probability (see Proposition 14 in Clémençon et al. (2017)), the empirical counterpart $\widehat{P}_N$ of a (strictly) stochastically transitive ranking distribution P can be of course non (S)ST. We propose below a trimming strategy based on the empirical ranking depth to recover a close (S)ST empirical ranking distribution and overcome this issue.

Algorithm 1: Ranking Depth Trimming

Input : Ranking dataset $\mathcal{D}_N = \{\Sigma_1, \ldots \Sigma_N\}$ and distribution $\widehat{P}_N = (1/N) \sum_{i=1}^N \delta_{\Sigma_i}$.

Output : Dataset $\mathcal{D} \subset \mathcal{D}_N$ of size $N_{\mathcal{D}} \leq N$ and (S)ST ranking distribution $\widehat{P}_{\mathcal{D}} = (1/N_{\mathcal{D}}) \sum_{\sigma \in \mathcal{D}} \delta_\sigma$

- Initialize: $\mathcal{D} = \mathcal{D}_N$;

while $\hat{P}_{\mathcal{D}}$ *is not (S)ST* **do**

- Determine the least deep rankings in $\mathcal{D}$: $\mathcal{O}_{\mathcal{D}} := \arg\min_{\sigma \in \mathcal{D}} D_{\hat{P}_N}(\sigma)$;
- Update the ranking dataset $\mathcal{D} \setminus \mathcal{O}_{\mathcal{D}} \to \mathcal{D}$

Based on the ranking dataset $\mathcal{D}$ output by Algorithm 1, a (S)ST empirical distribution $\widehat{P}_{\mathcal{D}}$ can be computed, whose Kemeny medians are obtained in a straightforward manner, *cf* Proposition 6, avoiding the search of solutions of a NP-hard minimization problem of type (1), see Hudry (2008). As empirically supported by the experiments displayed in the next section, this procedure allows for a fast, accurate and robust recovery of consensus rankings. Indeed, the time complexity of Algorithm 1 is in $n\log(n)N^2\eta$, where n is the number of items, N the number of samples ($n\log(n)$ is the complexity of computing Kendall's-τ for a pair of data using e.g. Merge Sort algorithm and N^2 to recompute the expected value of Kendall's-τ to the whole dataset for every point of the dataset) and η is the (unknown) number of iterations required to obtain a SST dataset from a non-SST one.

5 Applications - Experiments

In order to illustrate the relevance of ranking depth notion, we now show that it can be used to perform a wide variety of tasks in the statistical analysis of ranking data, including those listed below:

- Fast and robust consensus ranking
- Ranking data visualization
- Detection of outlying rankings
- The two-sample (homogeneity) problem in $\mathfrak{S}_n$.

Further experimental results on real ranking data are provided and discussed in the Supplementary Material, and the code is available here: github.com/RankingDepth/Ranking_depth_function.

5.1 Fast/Robust Consensus Ranking

The trimming strategy proposed in sec 4.2 shows that we can recover smooth SST distributions from any empirical data, and perform ranking aggregation by simply identifying the deepest ranking: this procedure is fast, straightforward, and robust, in the sense that we can recover accurate medians even in contaminated settings. We support this claim by both experiments and a theoretical proposition below.

We consider a dataset drawn from a "clean" distribution P (10000 points drawn from a Mallows distribution with $n = 12$ items, center σ_0 and $\phi = 0.90$) that has been contaminated by rankings from another distribution (2000 points drawn from a Mallows distribution with opposite center and $\phi = 0.40$). We use the trimming strategy described in algorithm 1 to remove rankings until the empirical distribution becomes SST and thus considered clean once again. We show in Figure 1 the depth of clean (blue) and adversarial (red) rankings before trimming (a) and after trimming (b),

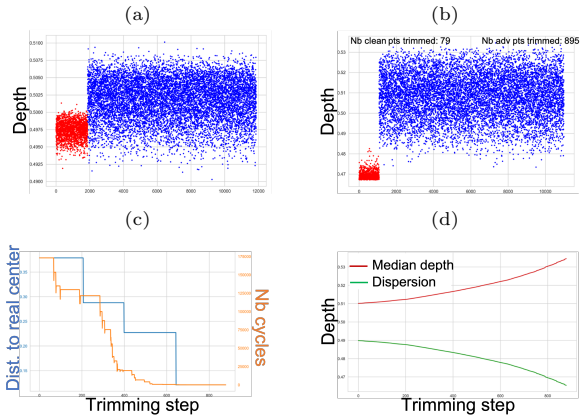


Figure 1: Depth plots before (a) and after (b) trimming with adversarial (red) and clean (blue) points; evolution of candidate median (deepest ranking) distance to real median and number of cycles through trimming (c); evolution of median depth and sample dispersion through trimming (d).

the performance of the median computed at each step of the trimming procedure evaluated as its Kendall τ distance to the real center of the clean Mallows distribution (c), and the depth of the median during the trimming procedure (d). The depth function is able to identify mainly adversarial rankings and remove them during the trimming procedure, which conducts to a cleaner dataset after the procedure and a far more accurate median σ^*.

Mechanical Turk Dots dataset. We show the robustness of depth-based medians on a real dataset where participants ranked point clouds according to their size (Mao et al., 2013). A ground truth ranking exists, and we contaminated 1/4 of the dataset by swapping random rankings before trimming: figure 2 (b) shows that we indeed recovered the ground truth ranking after the trimming strategy even if contaminated rankings were not obviously different from clean one (fig. 2 (a)).

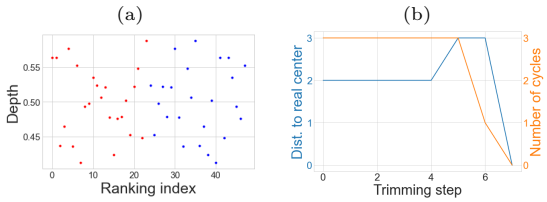


Figure 2: Depth plots before trimming with swapped (red) and clean (blue) points; evolution of candidate median (deepest ranking) distance to real median and number of cycles through trimming (b)

Theoretical robustness result. We derive specific robustness results when using depth-based trimming by emulating the classical notion of breakdown point (see Donoho and Gasko (1992)). Let us consider the classical Borda estimator (which orders the items based on the score $B(i) = \sum_{\sigma \in S_N} \sigma(i)$, see Dwork et al. (2001); Fligner and Verducci (1988); Caragiannis et al. (2013); Collas and Irurozki (2021)) and a *depth-trimmed* Borda estimator based on the scores $B_\mu(i) = \sum_{\sigma \in S_N} w(\sigma)\sigma(i)$, where $w(\sigma) = \mathbb{I}(D_N(\sigma) > \mu)$ (only the rankings with depth higher than μ are kept). Let σ_S^{B} (resp. $\sigma_S^{\text{DT-B}}$) be the Borda (resp. depth-trimmed Borda) estimator of a sample S. The Borda estimator is said to be δ-broken for sample size N and for a distribution P if for any sample $S_N \sim P$ of size N, there exists an adversarial sample A such that $d_\tau(\sigma^{\mathrm{B}}_{S_N}, \sigma^{\mathrm{B}}_{S_N \cup A}) \geq \delta$. The smallest cardinality of the adversarial sample A such that the estimator is δ-broken for size $N \to \infty$ is called here the *δ-breakdown points* of the estimator on distribution P, and we write $\epsilon_\delta^{\mathrm{B}}(P)$ (resp. $\epsilon_\delta^{\text{DT-B}}(P)$) such statistic for the Borda (resp. depth-trimmed Borda) estimator. Breakdown points measure the robustness of an estimator on a given distribution: we state that the classical Borda estimator is less robust than the depth-trimmed one on generic distributions.

Proposition 8. *Let μ be the trimming threshold and P a distribution such that $\mathbb{E}_P[D_P(\Sigma)] > \mu$. Let $\sigma^* = \arg\max_{\sigma \in \mathfrak{S}_n} D_P(\sigma)$ be the deepest ranking and $\pi = \arg\max_{\sigma | d_\tau(\sigma^*, \sigma) = \delta} D(\sigma)$ the ranking with highest depth among those at distance δ from the deepest ranking σ^*. Then, the breakdown points for Borda and depth-trimmed-Borda on P are related as follows,*

$$\frac{\epsilon_\delta^{B}(P)}{\epsilon_\delta^{DT\text{-}B}(P)} < \frac{D_P(\pi)}{\mu} < 1. \tag{2}$$

The proof, as well as more results on the robustness of Borda estimators, are provided in section B.4 of the supplementary.

5.2 Graphical Methods and Visual Inference

The analysis of rankings suffers from the lack of graphical displays and diagrams, such as *probability plots* or *histograms*, for gaining insight into the structure of the data. Ranking depths can be readily used to design a visual diagnostic tool for ranking data, extending the Depth *vs.* Depth plot (*DD*-plot in abbreviated form) were originally introduced by Liu et al. (1999) for multivariate data. For two samples of rankings $\mathbf{\Sigma^1} = \{\sigma_1^1, \ldots, \sigma_{N_1}^1\}$ and $\mathbf{\Sigma^2} = \{\sigma_1^2, \ldots, \sigma_{N_2}^2\}$, with corresponding empirical measures $\widehat{P^1}_{N_1}$ and $\widehat{P^2}_{N_2}$, the ranking *DD*-plot is obtained by plotting in the Euclidean plane the points:

$$\left\{ \left(D_{\widehat{P^1}_{N_1}}(\sigma), D_{\widehat{P^2}_{N_2}}(\sigma) \right) : \sigma \in \mathbf{\Sigma^1} \cup \mathbf{\Sigma^2} \right\}. \tag{3}$$

Position	$d_\tau(\sigma_1^*, \sigma_2^*)$	ϕ_1	ϕ_2	N_1	N_2
(a)	15	e^{-1}	e^{-1}	250	250
(b)	0	$\mathrm{e}^{-0.5}$	e^{-2}	250	250
(c)	15	$\mathrm{e}^{-0.5}$	e^{-2}	250	250
(d)	15	$\mathrm{e}^{-0.5}$	e^{-2}	400	100

Table 1: Parameters for pairs of samples drawn from Mallows-Kendall distribution used for Figure 3.

Depending on the distance d chosen, such a plot allows to reflect location and scatter of two distributions on $\mathfrak{S}_n$, and their mutual position. To illustrate its diagnostic capacity, we plot in Figure 3 the ranking *DD*-plots relative to the Kendall τ distance and four pairs of samples stemming from Mallows distribution with parameters defined in Table 1. (In this and subsequent figures the depth is re-scaled to $[0, 1]$ by diving by $\|d\|_\infty$.) A few remarks can be made: For distributions differing in: 1) location only (a), the ranking *DD*-plot is symmetric w.r.t. the diagonal, 2) scatter only (b), observations from one distribution will be attributed systematically higher depth values, 3) both location and scatter (c), they can be distinguished and 4) number of the observations, it does not influence the general picture (d).

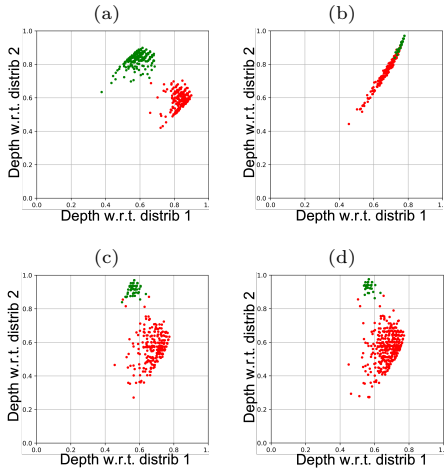


Figure 3: Ranking *DD*-plot corresponding to Mallows distributions with parameters described in Table 1.

5.3 Outlier Detection in Ranking Data

We now place ourselves in the situation where a single sample of rankings is observed. For simplicity, we consider the case where the underlying ranking distribution

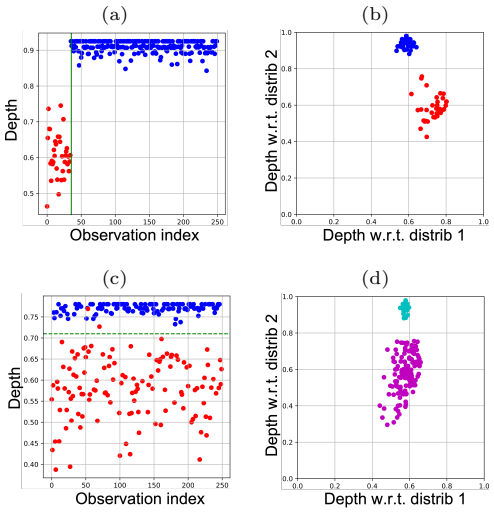


Figure 4: Depth plots (a,c) and DD-plots (b,d) for a mixture of Mallows-Kendall distributions. (a)-(b): distant centers and different size for the two components of the mixture. (c)-(d): closer centers and same size.

is an unbalanced mixture of two Mallows distributions (for $n = 10$), strongly differing in size ($N_1 = 35$ and $N_2 = 215$), with distant centers ($d_\tau(\sigma_1^*, \sigma_2^*) = 15$) and parameters $\phi_1 = \mathrm{e}^{-0.5}$ and $\phi_2 = \mathrm{e}^{-2.5}$. Figure 4 (a) shows the ranking depth (relative to Kendall τ) of each observation computed w.r.t. to the entire sample. We observe, that despite the unavailability of labels, the ranking depth clearly distinguishes the two different components. It thus permits to perform a typical anomaly detection task in the context of ranking data, where the differing minority of permutations are viewed as abnormal rankings. The diagnostic ranking DD-plot (b) based on the identified information about the components confirms the differences.

Consider next the case of a mixture with closer centers ($d_\tau(\sigma_1^*, \sigma_2^*) = 11$) and equal sizes ($N_1 = N_2 = 125$), with parameters $\phi_1 = \mathrm{e}^{-0.25}$ and $\phi_2 = \mathrm{e}^{-2.5}$. The depth plot (c) w.r.t. to the entire sample reflects how easily we can cluster the ranking dataset into two components (we deliberately shuffle the indices and keep colors for illustrative purposes), and we suggests a separating threshold (on the level of depth = 0.71), which in this particular case allows for two mistaking assignments. For the diagnostic ranking DD-plot (d), we honestly include this mistake, and change the colors to underline this impurity.

5.4 Rankings - Homogeneity Testing

Depth can further be used to provide a formal inference, which we exemplify as a nonparametric test of homogeneity between two Plackett-Luce distributions

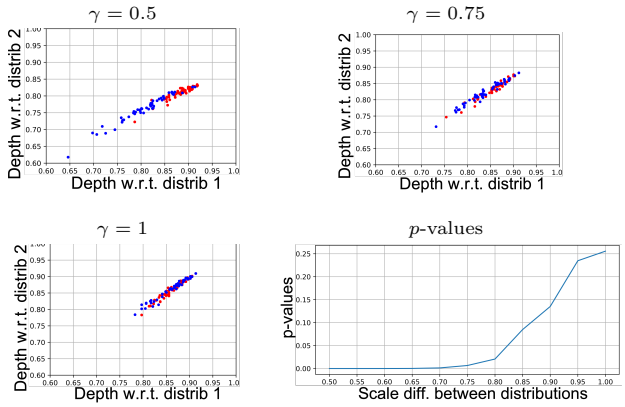


Figure 5: DD-plots of a pair of P-L distributions with gradually decreasing difference between them based on parameter γ and the corresponding average p-values for the test of homogeneity.

(Critchlow et al., 1991) with $n = 10$. The first one (red in Figure 5) is generated using the parameters $\boldsymbol{w}_1 = (\mathrm{e}^9, ..., \mathrm{e}^0)$, the second one represents its changed version $\boldsymbol{w}_2 = (\mathrm{e}^{\gamma 9}, ..., \mathrm{e}^{\gamma 0})$. We gradually increase γ from 0.5 (substantial difference) to 1 (equal in distribution), and provide the p-values of the Wilcoxon rank-sum test averaged over 100 repetitions in Figure 5. The test is performed using the reference sample (of size 500) from the first distribution, with tested sample sizes being equal (= 50) for both distributions (see Lafaye De Micheaux et al. (2020) for details on the testing procedure and Liu and Singh (1993) for more details). Figure 5 shows how the p-values detect very well the difference between the two distributions when it is the case, giving a formal inference to the ranking DD-plot visualization, whereas, remarkably, the (parametric) nature of the underlying ranking models is not used at all by the procedure. We also underline that, in a similar fashion, ranking depth-based *goodness-of-fit* statistics could be computed, in order to evaluate how well a specific ranking model fits a ranking dataset.

Student dataset. We now explore our homogeneity testing machinery on a real dataset (available at `https://github.com/ekhiru/students-dataset`) composed of rankings from students (with a ground truth answer) before (red) and after (blue) taking the related course. The diagnostic DD-plot of the two cohorts together with p-values over 1000 random repetitions and the asymptotic density under H_0 are indicated in Figure 6: they illustrate the improvement of the students' knowledge after the class.

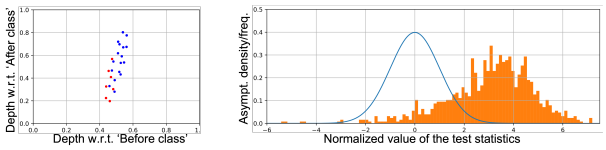


Figure 6: Left: *DD*-plot for 'before class' (red) and 'after class' (blue) students. Right: p-values of the homogeneity test.

Conclusion

In this paper, we have extended the concept of statistical depth to ranking data, in order to apply the notions of quantiles, order statistics and ranks to the latter, overcoming hence the lack of natural order and vector space structure on $\mathfrak{S}_n$. We have listed the desirable properties a ranking depth should satisfy to emulate these notions appropriately and shown that the same metric approach as that, widely used, to deal with ranking aggregation, permits to build depth functions on $\mathfrak{S}_n$ that fulfill them in many situations. Theoretical results proving that ranking depths and related quantities can be accurately estimated by their empirical versions with guarantees have been established. We have also shown that the methodology promoted can be successfully applied to a wide variety of problems, ranging from fast and robust consensus ranking to the design of ranking data visualization techniques through the detection of outlying rankings. Both the theoretical and empirical results are very encouraging and paves the way to a more systematic use of the ranking depth concept for the statistical analysis of ranking data.

References

Alvo, M. and Yu, P. L. H. (2014). *Statistical Methods for Ranking Data*. Springer-Verlag, New York.

Bartholdi III, J. J., Tovey, C. A., and Trick, M. A. (1989). The computational difficulty of manipulating an election. *Social Choice and Welfare*, 6(3):227–241.

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Brunel, V. E. (2019). Concentration of the empirical level sets of tukey's halfspace depth. *Probability Theory and Relative Fields*, 173:1165–1196.

Burr, M. A. and Fabrizio, R. J. (2017). Uniform convergence rates for halfspace depth. *Statistics and Probability Letters*, 124:33–40.

Busa-Fekete, R., Fotakis, D., Szörényi, B., and Zampetakis, M. (2019). Optimal Learning of Mallows Block Model. In *Conference on Learning Theory (COLT)*.

Busa-Fekete, R., Hüllermeier, E., and Szörényi, B. (2014). Preference-based rank elicitation using statistical models: the case of Mallows. In *Proceedings of International Conference on Machine Learning (ICML) 2014*, pages 1071–1079.

Caragiannis, I., Procaccia, A. D., and Shah, N. (2013). When do noisy votes reveal the truth? In *Proceedings of the Fourteenth ACM Conference on Electronic Commerce*, pages 143–160, New York. ACM.

Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association*, 91(434):862–872.

Chernozhukov, V., Galichon, A., Hallin, M., and Henry, M. (2017). Monge–kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1):223–256.

Clémençon, S., Korba, A., and Sibony, E. (2017). A learning theory of ranking aggregation. In *Proceeding of AISTATS 2017*.

Collas, F. and Irurozki, E. (2021). Concentric mixtures of Mallows models for top-k rankings: sampling and identifiability. In *International Conference on Machine Learning (ICML)*.

Condorcet, N. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. L'Imprimerie Royale, Paris.

Critchlow, D. E., Fligner, M. A., and Verducci, J. S. (1991). Probability models on rankings. *Journal of Mathematical Psychology*, 35(3):294–318.

Davenport, A. and Lovell, D. (2005). Ranking pilots in aerobatic flight competitions. Technical report, IBM Research Report RC23631 (W0506-079), TJ Watson Research Center, NY.

Davidson, D. and Marschak, J. (1959). Experimental tests of a stochastic decision theory. In Churchman, C. W. and Ratoosh, P., editors, *Measurement: Definitions and Theories*, pages 233–269. John Wiley.

de Borda, J.-C. (1781). Mémoire sur les élections au scrutin.

Desarkar, M. S., Sarkar, S., and Mitra, P. (2016). Preference relations based unsupervised rank aggregation for metasearch. *Expert Systems with Applications*, 49:86–98.

Deza, M. and Huang, T. (1998). Metrics on permutations, a survey. *Journal of Combinatorics, Information and System Sciences*.

Diaconis, P. and Graham, R. L. (1977). Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(2):262–268.

Donoho, D. L. and Gasko, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, 20:1803–1827.

Dwork, C., Kumar, R., Naor, M., and Sivakumar, D. (2001). Rank aggregation methods for the Web. In *International Conference on World Wide Web*, pages 613–622, New York. ACM.

Dyckerhoff, R. (2004). Data depths satisfying the projection property. *Allgemeines Statistisches Archiv*, 88(2):163–190.

Dyckerhoff, R., Ley, C., and Paindaveine, D. (2015). Depth-based runs tests for bivariate central symmetry. *Annals of the Institute of Statistical Mathematics*, 67(5):917–941.

Dyckerhoff, R. and Mozharovskyi, P. (2016). Exact computation of the halfspace depth. *Computational Statistics and Data Analysis*, 98:19–30.

Fishburn, P. C. (1973). Binary choice probabilities: on the varieties of stochastic transitivity. *Journal of Mathematical psychology*, 10(4):327–352.

Fligner, M. A. and Verducci, J. S. (1986). Distance based ranking models. *Journal of the Royal Statistical Society*, 48(3):359–369.

Fligner, M. A. and Verducci, J. S. (1988). Multistage ranking models. *Journal of the American Statistical Association*, 83(403):892–901.

Huang, T.-K., Weng, R. C., and Lin, C.-J. (2006). Generalized bradley-terry models and multi-class probability estimates. *The Journal of Machine Learning Research*, 7:85–115.

Hudry, O. (2008). NP-hardness results for the aggregation of linear orders into median orders. *Annals of Operations Research*, 163:63–88.

Irurozki, E., Calvo, B., and Lozano, J. (2019a). Mallows and generalized Mallows model for matchings. *Bernoulli*, 25(2).

Irurozki, E., Calvo, B., and Lozano, J. A. (2019b). PerMallows: An R package for mallows and generalized mallows models. *Journal of Statistical Software*, 71.

Jiao, Y., Korba, A., and Sibony, E. (2016). Controlling the distance to a kemeny consensus without computing it. In *Proceeding of International Conference on Machine Learning (ICML) 2016*.

Jörnsten, R. (2004). Clustering and classification based on the l_1 data depth. *Journal of Multivariate Analysis*, 90(1):67–89.

Kamishima, T. (2013). Nantonac collaborative filtering: Recommendation based on order responses. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 583–588, New York. Association for Computing Machinery.

Koshevoy, G. and Mosler, K. (1997). Zonoid trimming for multivariate distributions. *The Annals of Statistics*, 25(5):1998–2017.

Lafaye De Micheaux, P., Mozharovskyi, P., and Vimond, M. (2020). Depth for curve data and applications. *Journal of the American Statistical Association*. in press.

Lange, T., Mosler, K., and Mozharovskyi, P. (2014). Fast nonparametric classification based on data depth. *Statistical Papers*, 55(1):49–69.

Lebanon, G. and Lafferty, J. (2002). Cranking: Combining rankings using conditional probability models on permutations. In *Proceedings of the 19th International Conference on Machine Learning*, pages 363–370.

Liu (1990). On a notion of data depth based upon random simplices. *The Annals of Statistics*, 18(1):405–414.

Liu, A. and Moitra, A. (2018). Efficiently learning mixtures of mallows models. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 627–638. IEEE.

Liu, A., Zhao, Z., Liao, C., Lu, P., and Xia, L. (2019a). Learning Plackett-Luce Mixtures from Partial Preferences. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Liu, R. Y., Parelius, J. M., and Singh, K. (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference. *The Annals of Statistics*, 27(3):783–858. With discussion and a rejoinder by Liu and Singh.

Liu, R. Y. and Singh, K. (1993). A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association*, 88(421):252–260.

Liu, X., Mosler, K., and Mozharovskyi, P. (2019b). Fast computation of tukey trimmed regions and median in dimension $p > 2$. *Journal of Computational and Graphical Statistics*, 28(3):682–697.

Lu, T. and Boutilier, C. (2014). Effective Sampling and Learning for Mallows Models with Pairwise-Preference Data. *Journal of Machine Learning Research*.

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. John Wiley and Sons, New York.

Ma, Y., Genton, M. G., and Parzen, E. (2011). Asymptotic properties of sample quantiles of discrete distributions. *Annals of the Institute of Statistical Mathematics*, 63(2):227–243.

Mallows, C. L. (1957). Non-null ranking models. *Biometrika*, 44(1-2):114–130.

Mao, A., Procaccia, A., and Chen, Y. (2013). Better human computation through principled voting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27.

Mosler, K. (2013). Depth statistics. In Becker, C., Fried, R., and Kuhnt, S., editors, *Robustness and Complex Data Structures: Festschrift in Honour of Ursula Gather*, pages 17–34. Springer.

Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statistics and Probability Letters*, 1(6):327–332.

Patel, T., Telesca, D., Rallo, R., George, S., Xia, T., and Nel, A. E. (2013). Hierarchical rank aggregation with applications to nnanotoxicology. *Journal of Agricultural, Biological, and Environmental Statistics*, 18(2):159–177.

Plackett, R. L. (1975). The analysis of permutations. *Applied Statistics*, 2(24):193–202.

Pokotylo, O., Mozharovskyi, P., and Dyckerhoff, R. (2019). Depth and depth-based classification with R-package ddalpha. *Journal of Statistical Software, Articles*, 91(5):1–46.

Serfling, R. (2006). Depth functions in nonparametric multivariate inference. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 72.

Shah, N. B., Balakrishnan, S., Guntuboyina, A., and Wainright, M. J. (2015). Stochastically transitive models for pairwise comparisons: statistical and computational issues. arXiv preprint `arXiv:1510.05610`.

Shah, N. B. and Wainwright, M. J. (2015). Simple, robust and optimal ranking from pairwise comparisons.

Sheather, S. J. and Marron, J. S. (1990). Kernel quantile estimators. *Journal of the American Statistical Association*, 85(410):410–416.

Tukey, J. W. (1975). Mathematics and the picturing of data. In James, R. D., editor, *Proceedings of the International Congress of Mathematicians*, volume 2, pages 523–531. Canadian Mathematical Congress.

Vardi, Y. and Zhang, C.-H. (2000). The multivariate l_1-median and associated data depth. *Proceedings of the National Academy of Sciences*, 97(4):1423–1426.

Vitelli, V., Sørensen, Ø., Crispino, M., Frigessi, A., and Arjas, E. (2018). Probabilistic preference learning with the Mallows rank model. *Journal of Machine Learning Research*, 18(1).

Zhao, Z. and Xia, L. (2019). Learning Mixtures of Plackett-Luce Models from Structured Partial Orders. In *Advances in Neural Information Processing Systems*, pages 10143–10153.

Zuo, Y. and Serfling, R. (2000a). General notions of statistical depth function. *The Annals of Statistics*, 28(2):461–482.

Zuo, Y. and Serfling, R. (2000b). Structural properties and convergence results for contours of sample statistical depth functions. *The Annals of Statistics*, 28(2):483–499.

Supplementary Material

Contents

A Ranking Distributions - Popular Examples **12**

B Technical Proofs **14**

B.1 Conditions for satisfying the desirable properties . . . 14

B.1.1 Proof of Proposition 1 (invariance) . . . 14

B.1.2 Proof of Proposition 2 (maximality at the center) . . . 14

B.1.3 Proofs of Propositions 3 and 4 (monotonicity) . . . 18

B.2 Proof Proposition 7 (learning rate bounds) . . . 20

B.3 Proofs of Propositions 5, 6 and 15 (results for Kendall τ - Mallows model) . . . 21

B.4 Proof of Proposition 8 (Borda estimators' robustness) . . . 22

C Further results **25**

C.1 Ranking quantile function . . . 25

C.2 Pairwise comparisons as an alternative statistical framework . . . 25

D Additional experiments **26**

D.1 Trimming strategy . . . 26

D.2 Visual analysis . . . 27

D.3 Application to real data . . . 29

D.3.1 Student dataset . . . 29

D.3.2 Sushi dataset . . . 29

D.3.3 Mechanical Turk Dots dataset . . . 30

D.3.4 Netflix Prize dataset . . . 31

A Ranking Distributions - Popular Examples

Below we recall some popular ranking models. They will be next used to illustrate some of the properties involved in the theoretical analysis carried out.

Proposition 9. *The symmetry center property for rankings has two versions, a weak and a strong one, see Critchlow et al. (1991).*

1. *Strong unimodality: ranking model P is said to be strongly unimodal iff there exists a modal ranking σ^* such that for every pair i, j such that $\sigma^*(i) < \sigma(j)$ and any permutations σ such that $\sigma(i) = \sigma(j) - 1$ then $P(\sigma) \geq P(\sigma\tau_{ij})$, where $\sigma\tau_{ij}(i) = \sigma(j)$, $\sigma\tau_{ij}(j) = \sigma(i)$ and $\sigma\tau_{ij}(k) = \sigma(k)$ for $k \neq i, j$.*

2. *Complete consensus: ranking model P is said to have complete consensus iff there exists a modal ranking σ^* such that for every pair i, j such that $\sigma^*(i) < \sigma(j)$ and any permutations σ such that $\sigma(i) < \sigma(j)$ then $P(\sigma) \geq P(\sigma\tau_{ij})$, where $\sigma\tau_{ij}(i) = \sigma(j)$, $\sigma\tau_{ij}(j) = \sigma(i)$ and $\sigma\tau_{ij}(k) = \sigma(k)$ for $k \neq i, j$. Complete consensus implies strong unimodality.*

Example 1. (MALLOWS DISTRIBUTION) Taking $d = d_\tau$, the Mallows model introduced in Mallows (1957) is the unimodal distribution P_θ on $\mathfrak{S}_n$ parametrized by $\theta = (\sigma_0, \phi_0) \in \mathfrak{S}_n \times (0, 1]$: $\forall \sigma \in \mathfrak{S}_n$,

$$P_\theta(\sigma) = (1/Z_0) \exp(d_\tau(\sigma_0, \sigma) \log \phi_0), \tag{4}$$

where $Z_0 = \sum_{\sigma \in \mathfrak{S}_n} \exp(d_\tau(\sigma_0, \sigma) \log \phi_0)$ is a normalization constant. One may easily show that Z_0 is independent from σ_0 and that $Z_0 = \prod_{i=1}^{n-1} \sum_{j=0}^{i} \phi_0^j$. When $\phi_0 < 1$, the permutation σ_0 of reference is the mode of distribution P_{θ_0}, as well as its unique median relative to d_τ. Observe in addition that the smallest the parameter ϕ_0, the spikiest the distribution P_{θ_0}. In contrast, P_{θ_0} is the uniform distribution on $\mathfrak{S}_n$ when $\phi_0 = 1$. As explained in section 3, ranking depth functions relative to the Kendall τ distance can be expressed as a function of the pairwise probabilities $p_{i,j} = \mathbb{P}\{\Sigma(i) < \Sigma(j)\}$, $1 \leq i \neq j \leq n$. Notice also that $||d_\tau||_\infty = \binom{n}{2}$. Consider again the Mallows model P_θ recalled in Example 1. In this case, a closed-from expression of the $p_{i,j}$'s is available, see *e.g.* Theorem 2 in Busa-Fekete et al. (2014). Setting $h(k, \phi_0) = k/(1 - \phi_0^k)$ for $k \geq 1$, one can then show that the ranking depth function relative to P_θ and d_τ is: $\forall \sigma \in \mathfrak{S}_n$, $D_{P_\theta}(\sigma) = \binom{n}{2} - \sum_{\sigma(i) > \sigma(j)} H(\sigma_0(j) - \sigma_0(i),\ \phi_0)$, where $H(k, \phi_0) = h(k+1, \phi_0) - h(k, \phi_0)$ and $H(-k, \phi_0) = 1 - H(k, \phi_0)$ for $k \geq 1$. Mallows is adapted naturally to work with extensions of rankings, such as from pairwise preferences Lu and Boutilier (2014), and partial rankings Vitelli et al. (2018)

Mallows satisfies the complete consensus property, see Property 9, when $\theta < 1$.

The most popular extensions in the literature are Generalized Mallows models Fligner and Verducci (1986), Irurozki et al. (2019a) and Mallows Block models Busa-Fekete et al. (2019). They define different dispersion parameters for different ranking positions to model distributions in which there is high certainty in the top-ranked items and uncertainty at the bottom. These models still satisfy the complete consensus property, see Property 9, when $\theta < 1$.

We also point out that model (4) can be extended in a straightforward manner, by considering alternative distances d, including those described in Section 2.2 and other right invariant distances such as Cayley and Ulam, all of which satisfy the complete consensus property, see Property 9, when $\theta < 1$.

The maximality at center is broken in more general ranking distributions with the form of mixtures of Mallows models. Mixtures have been studied in practical and theoretical settings, see *e.g.* Lebanon and Lafferty (2002); Liu and Moitra (2018); Collas and Irurozki (2021).

Example 2. (PLACKETT-LUCE (PL) DISTRIBUTION) PL assumes that rankings are generated in a stage wise manner: the most preferred item is chosen first, then the second preferred one, ... There is independence among stages, that is, the probability of an item being chosen at a particular stage is only proportional to the remaining items at this stage and independent of the order of the items that have already been chosen. Thus, PL is parametrized by $\boldsymbol{v} \in \mathbb{R}^n$, where $v(i)$ is proportional to the probability of choosing item i as the preferred item at any stage (among the remaining ones). The probability of each ranking is given as

$$P_v(\sigma) = \prod_{i=1}^{n} \frac{\sigma^{-1}(i)}{\sum_{j=i}^{n} \sigma^{-1}(j)}. \tag{5}$$

The median ranking is the permutation that orders the weights decreasingly. The pairwise probabilities of items i and j have a closed-form expression involving only the weights of bot items, $p_{i,j} = \frac{v_i}{v_i + v_j}$. PL's stage wise ranking process implies that adaptation to top-k and rankings is natural Liu et al. (2019a).

The PL models satisfy the complete consensus property, see Property 9 for every distribution other than the uniform. Clearly, the maximality at center does not hold for mixtures of PL in general. Note that there is a body of research on PL mixtures Liu et al. (2019a); Zhao and Xia (2019).

Example 3. (MALLOWS-BRADLEY-TERRY DISTRIBUTION) Mallows-Bradley-Terry is a ranking model induced by paired comparisons in which the pairwise probability of items i and j have the form

$$p_{i,j} = \frac{v_i}{v_i + v_j},$$

where v_i is the parameter associated to item i for $\boldsymbol{v} \in \mathbb{R}^n$. The probability of ranking σ is then

$$p(\sigma) = Z(\boldsymbol{v}) \prod_{i=1}^{n-1} (v_{\sigma^{-1}(i)})^{n-i},$$

where Z is a normalization constant. See Huang et al. (2006) for generalizations.

Example 4. (PAIRWISE DISTRIBUTIONS) All the above models can be written as a $n \times n$ matrix of pairwise probabilities $p_{i,j}$ (describing the probability of item i being preferred to item j) with restricted forms of its entries. Each of the models imposes different restrictions in the entries of the pairwise probabilities $p_{i,j}$ but one could consider arbitrary values. We next lines characterize the properties of models with arbitrary entries $p_{i,j}$.

- P is strongly unimodal if and only if its entries are weakly stochastically transitive for some reordering of the rows an columns, as defined in Proposition 5.
- P has complete consensus if and only if its entries are strongly stochastically transitive for some reordering of the rows an columns. A probability distribution P on $\mathfrak{S}_n$ is said to be strongly stochastically transitive iff, for all $(i,j,k) \in [\![n]\!]^3$, we have: $p_{i,j} \geq 1/2$ and $p_{j,k} \geq 1/2 \Rightarrow p_{i,k} \geq \max\{p_{i,j}, p_{j,k}\}$ and $p_{i,j} \neq 1/2$ for all $i < j$.

B Technical Proofs

B.1 Conditions for satisfying the desirable properties

B.1.1 Proof of Proposition 1 (invariance)

We elaborate now on the invariance property 1. Recall that a distance is right invariant iff for every triplet of permutations $(\sigma, \pi, \nu) \in \mathfrak{S}_n d(\sigma, \nu) = d(\sigma\pi, \nu\pi)$. Finally, the inverse of permutation σ is denoted by σ^{-1}.

Let us first recall the invariance property for distributions and for depths 1 and our proposition 1:

Property 1. (INVARIANCE) *For any $\pi \in \mathfrak{S}_n$, consider the ranking distribution πP defined by: $(\pi P)(\sigma) = P(\sigma\pi^{-1})$ for all $\sigma \in \mathfrak{S}_n$. It holds that: $D_P(\sigma) = D_{\pi P}(\sigma\pi)$ for all $(\sigma,\ \pi) \in \mathfrak{S}_n^2$.*

Proposition 1. (INVARIANCE) *Suppose that d is right-invariant, i.e. $d(\nu\pi, \sigma\pi) = d(\nu, \sigma)$ for all $(\nu, \pi, \sigma) \in \mathfrak{S}_n^3$, the ranking depth $D_P^{(d)}$ satisfies the Property 1.*

Proof.

$$\begin{aligned} D_{\pi P}(\sigma\pi) =& \mathbb{E}_{\pi P}[||d||_\infty - d(\sigma\pi, \Sigma)] = ||d||_\infty - \sum_{\nu \in \mathfrak{S}_n} (\pi P)(\nu) d(\sigma\pi, \nu) \\ =& ||d||_\infty - \sum_{\nu \in \mathfrak{S}_n} P(\nu\pi^{-1}) d(\sigma\pi, \nu) = ||d||_\infty - \sum_{\nu' \in \mathfrak{S}_n} P(\nu'\pi\pi^{-1}) d(\sigma\pi, \nu'\pi) \\ =& ||d||_\infty - \sum_{\nu' \in \mathfrak{S}_n} P(\nu') d(\sigma, \nu') = D_P(\sigma). \end{aligned} \tag{6}$$

□

B.1.2 Proof of Proposition 2 (maximality at the center)

First, we study the relation between the depth and the probability of permutations which will be key for the results on the following sections.

Proposition 10. *Let P be a SST distribution whose Kemeny's median is σ^*, and $\sigma^*(a) < \sigma^*(b)$. Let σ be a ranking such that $\sigma(a) + 1 = \sigma(b)$ and let t_{ab} be a transposition, i.e., $t_{ab}(a) = b$, $t_{ab}(b) = a$ and $t_{ab}(k) = k$ for all $k \neq a, b$. Then,*

$$D(\sigma) > D(\sigma t).$$

Proof. First, note that the composition σt_{ab} exchanges the ranks of items a and b, so $d(\sigma, \sigma^*) = d(\sigma t, \sigma^*) - 1$. We can rewrite $D(\sigma)$ in the following way,

$$\begin{aligned} D(\sigma) =& \binom{n}{2} - \sum_{i<j} p_{i,j} \mathbb{I}\{\sigma(i) > \sigma(j)\} - \sum_{i<j} p_{j,i} \mathbb{I}\{\sigma(i) < \sigma(j)\} \\ =& \binom{n}{2} - \sum_{i<j \wedge i,j \neq a,b} p_{i,j} \mathbb{I}\{\sigma(i) > \sigma(j)\} - \sum_{i<j \wedge i,j \neq a,b} p_{j,i} \mathbb{I}\{\sigma(i) < \sigma(j)\} - \\ & - p_{a,b} \mathbb{I}\{\sigma(a) > \sigma(b)\} - p_{ba} \mathbb{I}\{\sigma(a) < \sigma(b)\} \\ =& \binom{n}{2} - \sum_{i<j \wedge i,j \neq a,b} p_{i,j} \mathbb{I}\{\sigma(i) > \sigma(j)\} - \sum_{i<j \wedge i,j \neq a,b} p_{j,i} \mathbb{I}\{\sigma(i) < \sigma(j)\} - p_{b,a}. \end{aligned} \tag{7}$$

Where the first equality is the given by proposition 5. In the second we split the sum for positions $i = a$ and $j = b$ in the latter term and the rest of the pairs in the previous terms. In the third one, we recall that by assumption $\sigma(a) = \sigma(b) - 1$ and therefore $\mathbb{I}\{\sigma(a) < \sigma(b)\} = 1$ and $\mathbb{I}\{\sigma(a) > \sigma(b)\} = 0$. We rewrite in a similar way $D(\sigma t_{ab})$. For this part, recall that $\sigma t_{ab}(a) = \sigma(b)$, $\sigma t_{ab}(b) = \sigma(a)$ and $\sigma t_{ab}(k) = k$ for all $k \neq a, b$.

$$\begin{aligned} D(\sigma t_{ab}) =& \binom{n}{2} - \sum_{i<j} p_{i,j} \mathbb{I}\{\sigma t_{ab}(i) > \sigma t_{ab}(j)\} - \sum_{i<j} p_{j,i} \mathbb{I}\{\sigma t_{ab}(i) < \sigma t_{ab}(j)\} \\ =& \binom{n}{2} - \sum_{i<j \wedge i,j \neq a,b} p_{i,j} \mathbb{I}\{\sigma t_{ab}(i) > \sigma t_{ab}(j)\} - \sum_{i<j \wedge i,j \neq a,b} p_{j,i} \mathbb{I}\{\sigma t_{ab}(i) < \sigma t_{ab}(j)\} \\ & - p_{a,b} \mathbb{I}\{\sigma t_{ab}(a) > \sigma t_{ab}(b)\} - p_{b,a} \mathbb{I}\{\sigma t_{ab}(a) < \sigma t_{ab}(b)\} \\ =& \binom{n}{2} - \sum_{i<j \wedge i,j \neq a,b} p_{i,j} \mathbb{I}\{\sigma(i) > \sigma(j)\} - \sum_{i<j \wedge i,j \neq a,b} p_{j,i} \mathbb{I}\{\sigma(i) < \sigma(j)\} - \\ & - p_{a,b} \mathbb{I}\{\sigma(a) < \sigma(b)\} - p_{ba} \mathbb{I}\{\sigma(a) > \sigma(b)\} \\ =& \binom{n}{2} - \sum_{i<j \wedge i,j \neq a,b} p_{i,j} \mathbb{I}\{\sigma(i) > \sigma(j)\} - \sum_{i<j \wedge i,j \neq a,b} p_{j,i} \mathbb{I}\{\sigma(i) < \sigma(j)\} - p_{a,b} \end{aligned} \tag{8}$$

Therefore, for any two rankings σ and σt_{ab} such that $D(\sigma) > D(\sigma t_{ab})$, the following holds,

$$\begin{aligned} & D(\sigma) > D(\sigma t_{ab}) \\ \Leftrightarrow & \binom{n}{2} - \sum_{i<j \wedge i,j \neq a,b} p_{i,j} \mathbb{I}\{\sigma(i) > \sigma(j)\} - \sum_{i<j \wedge i,j \neq a,b} p_{j,i} \mathbb{I}\{\sigma(i) < \sigma(j)\} - p_{b,a} \\ & > \binom{n}{2} - \sum_{i<j \wedge i,j \neq a,b} p_{i,j} \mathbb{I}\{\sigma(i) > \sigma(j)\} - \sum_{i<j \wedge i,j \neq a,b} p_{j,i} \mathbb{I}\{\sigma(i) < \sigma(j)\} - p_{a,b} \\ \Leftrightarrow & p_{b,a} < p_{a,b}. \end{aligned} \tag{9}$$

For any SST model P with whose median is σ^*, and where $\sigma^*(a) < \sigma^*(b)$, it holds (by definition) that $p_{b,a} < p_{a,b}$, which concludes the proof. □

Let us first recall Property 2 (maximality) and Proposition 2.

Property 2. (Maximality at center) *For any probability distribution P on $\mathfrak{S}_n$ that possesses a symmetry center σ_P (in a certain sense, e.g. w.r.t. to a given metric d on $\mathfrak{S}_n$), the depth function D_P takes its maximum value at it, i.e. $D_P(\sigma_P) = \max_{\sigma \in \mathfrak{S}_n} D_P(\sigma)$.*

Proposition 2. (Maximality at the center)*: The Spearman's footrule ranking depth satisfies Property 2 for any distribution P with a symmetry center. If P is SST in addition, then Kendall τ and Spearman ρ ranking depths satisfy Property 2 as well.*

We now discuss what is precisely meant by *center* in the ranking context. We derive two main definitions for a center:

- Following Tukey (1975); Zuo and Serfling (2000a), we emulate the notion of *half-space* symmetry (which is a very generic notion of symmetry) and define a notion of H-center, from which our proposition in the main paper stems from. Appart from our maximality proposition, we further provide results for distributions P having a H-center.

- We define a simpler notion of center based on a natural metric approach. We also provide maximality results based on this different notion of center, called a M-center.

H-center and maximality at center.

The following results (1) define a symmetry center inspired in the classical formulation of half-space symmetry Tukey (1975); Zuo and Serfling (2000a) and (2) shows that the Kendall's-τ, Spearman's footrule and Spearman ρ distances satisfy the maximality at center for the defined center.

Proposition 11. *Let us call "hyperplane" the sets $H_{i,j} = \{\sigma : \sigma(i) < \sigma(j)\}$, we define the H-center σ as $\sigma = \cap H_{i,j}$ for all $\{(i,j) : \sigma_0(i) < \sigma_0(j)\}$. For any P such that $p_{i,j} > p_{j,i}$ for all $\{(i,j) : \sigma_0(i) < \sigma_0(j)\}$ the H-center is σ_0.*

Proof. Firstly, we show that $P(\Sigma \in H_{i,j}) > P(\Sigma \in H_{j,i})$. This can be done by construction: For any ranking $\sigma \in H_{i,j}$ (for which $\sigma(i) = \sigma(j)$) we can construct $\sigma' \in H_{j,i}$ that swaps positions i and j. This construction defines a bijection between the rankings in both sets. The following relation holds: $p(\sigma') = p(\sigma)p_{j,i}/p_{i,j} < p(\sigma)$. Therefore, $P(\Sigma \in \cap H_{i,j}) > P(\Sigma \in \cap H_{j,i})$.

Secondly, it is clear that there is one and only one permutation in $\cap H_{i,j}$ and this is σ_0. We remark that its possible an H-center is defined (for this choice of P) by a smaller number of subsets, i.e., those $H_{i,j}$ for which $\sigma(i) = \sigma(j) - 1$. □

Proposition 12. *Let P be distribution for which there is an H-center both Kendall's-τ, Spearman's footrule and Spearman ρ based depths satisfy the maximality at center property for the H-center in Definition 11.*

Proof. As shown in Proposition 11, the H-center is σ_0 It remains to recall that Equation(9) in Proposition 10 states that for SST models and the Kendall's-τ distance $D(\sigma) > D(\sigma t_{i,j}) \Leftrightarrow p_{j,i} < p_{i,j}$.

For the Spearman's distance, let us show that $D_P(\sigma_0) \geq D_P(\sigma_1) \Leftrightarrow \mathbb{E}_P(d(\Sigma, \sigma_0)) \leq \mathbb{E}_P(d(\Sigma, \sigma_1))$, and the proof of our proposition will follow from direct application of this result.

Let σ be any permutation.

$$
\begin{aligned}
d(\sigma, \sigma_1) &= \sum_{k=1}^{N} |\sigma(k) - \sigma_1(k)| \\
&= \sum_{k \neq i,j} |\sigma(k) - \sigma_0(k)| + |\sigma(i) - \sigma_0(i) - 1| + |\sigma(j) - \sigma_0(j) + 1| \\
&= \begin{cases}
d(\sigma, \sigma_0) \text{ if } \sigma(i) < \sigma(j) \leq \sigma_0(i) < \sigma_0(j) \text{ or } \sigma_0(i) < \sigma_0(j) \leq \sigma(i) < \sigma(j) \\
\qquad \text{or } \sigma(j) < \sigma(i) \leq \sigma_0(i) < \sigma_0(j) \text{ or } \sigma_0(i) < \sigma_0(j) \leq \sigma(j) < \sigma(i) \\
d(\sigma, \sigma_0) + 2 \text{ if } \sigma(i) < \sigma_0(i) < \sigma_0(j) < \sigma(j) \\
d(\sigma, \sigma_0) - 2 \text{ if } \sigma(j) \leq \sigma_0(i) < \sigma_0(j) \leq \sigma(i)
\end{cases}
\end{aligned}
$$

Notice the use of color: in blue are cases where i and j are ranked by σ the same way as does σ_0, and in orange

are the opposite cases. Then

$$\begin{aligned}
&\mathbb{E}_P(d(\sigma,\sigma_0)) \leq \mathbb{E}_P(d(\sigma,\sigma_1)) \\
\Leftrightarrow &\sum_\sigma \left[\mathbb{I}(\text{blue cases}) - \mathbb{I}(\text{orange cases})\right] \mathbb{P}(\Sigma=\sigma) \geq 0 \\
\Leftrightarrow &p_{i,j} - (1-p_{i,j}) \geq 0 \\
\Leftrightarrow &p_{i,j} \geq 1/2
\end{aligned}$$

For the Spearman ρ case, the proof is similar. In this case, we have $d(\sigma,\sigma') = c - 2 < \sigma, \sigma' >$ where c is a constant. Then $\mathbb{E}_\Sigma(d(\Sigma,\sigma)) = c - 2 < \sigma, \mathbb{E}(\Sigma) >$. Thus,

$$\begin{aligned}
\mathbb{E}_\Sigma(d(\sigma_0,\Sigma)) \leq \mathbb{E}_\Sigma(d(\sigma_1,\Sigma)) &\Leftrightarrow < \sigma_0, \mathbb{E}_\Sigma > \geq < \sigma_1, \mathbb{E}_\Sigma > \\
&\Leftrightarrow \sum_{k=1}^n [\sigma_0(k) - \sigma_1(k)] \sum_\sigma \mathbb{P}(\Sigma=\sigma)\sigma(i) \geq 0 \\
&\Leftrightarrow \sum_\sigma \mathbb{P}(\Sigma=\sigma)(\sigma(j)-\sigma(i)) \quad \text{by def. of } \sigma_1 \text{ that swaps } (i,j) \text{ w.r.t } \sigma_0 \geq 0 \\
&\Leftrightarrow \mathbb{E}_\Sigma(\Sigma(j)-\Sigma(j)) \geq 0 \\
&\Leftrightarrow \mathbb{P}(\Sigma(j)-\Sigma(i)>0) = p_{i,j} \geq 1/2
\end{aligned}$$

This concludes the proof. □

M-center definition.

Let us focus on a more natural, metric-based center definition.

Definition 6. *σ_0 is M-center for distance d and distribution P if: $\forall(\sigma_1,\sigma_2,\sigma_3)$ such that $d(\sigma_0,\sigma_1) = d(\sigma_0,\sigma_2) < d(\sigma_0,\sigma_3)$, we have: $\mathbb{P}(\Sigma=\sigma_1) = \mathbb{P}(\Sigma=\sigma_2) \geq \mathbb{P}(\Sigma=\sigma_3)$.*

We have the following proposition:

Proposition 13. *If d is a symmetric distance, and if distribution P has a M-center for d, the maximality property is satisfied for distance d.*

Most distances (as the one studied in this paper) are symmetric. In addition, the proposition applies to Mallows models as they do exhibit a S-center.

Proof. Let σ_0 be a M-center for P and distance d, with (i,j) such that $\sigma_0(i) < \sigma_0(j) = \sigma_0(i)+1$. Let σ_1 be the same ranking as σ_0 except it swaps the ranks of i and j.

We show that $D_P(\sigma_0) > D_P(\sigma_0)$ i.e. $\mathbb{E}_P(d(\Sigma,\sigma_0)) < \mathbb{E}_P(d(\Sigma,\sigma_1))$ i.e. $\sum_\sigma \mathbb{P}(\Sigma=\sigma)\left[d(\sigma_1,\sigma) - d(\sigma_0,\sigma)\right] > 0$.

Let σ be any ranking such that $d(\sigma_0,\sigma) = d$. We have:

- (1) $d(\sigma_0,\sigma) < d(\sigma_1,\sigma) = d + c_{i,j}$ iff (i,j) is ranked the same way in σ_0 and σ
- (2) $d(\sigma_0,\sigma) > d(\sigma_1,\sigma) = d - c_{i,j}$ else,

where $c_{i,j} > 0$ is a constant depending only on (i,j). For example, if d is Kendall's tau, $c_{i,j} = 1$, if d is Spearman's footrule, $c_{i,j} = 2$, if d is Spearman's rho, $c_{i,j} = 2|\sigma(j) - \sigma(i)|$.

In addition, let us write $\#d$ the number of rankings at distance d from σ_0, which we can divide into the two groups (1) and (2). Let us then write $\#d(1)$ (resp. $\#d(2)$) the number of rankings σ at distance d from σ_0 that rank i and j the same way (resp. differently) as σ_0. We suppose the following: if $d \leq ||d||_\infty/2$, then $\#d(1) \geq \#d(2)$ (and if $d > ||d||_\infty/2$, then $\#d(1) \leq \#d(2)$), and more precisely, $|\#d(1) - \#d(2)| = k(d) = k(||d||_\infty - d)\ \forall\ d \leq ||d||_\infty/2$, meaning that this cardinality difference depends only on the distance to half of the maximal distance.

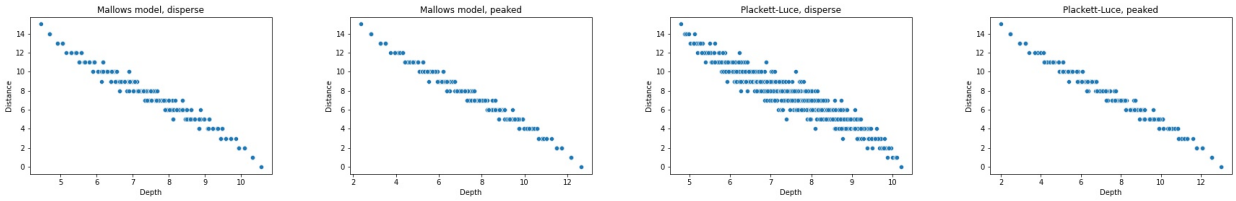
Let us also write $P_d = \mathbb{P}(\Sigma=\sigma)$ for any σ at distance d from σ_0.

$$
\begin{aligned}
\sum_{\sigma} \mathbb{P}(\Sigma = \sigma)\left[d(\sigma_1, \sigma) - d(\sigma_0, \sigma)\right] &= \sum_{d=0}^{||d||_\infty} P_d \times \#d \times |c_{i,j}| \\
&= \sum_{d=0}^{||d||_\infty} P_d \times (\#d(1) - \#d(2)) \times c_{i,j} \\
&= \sum_{d=0}^{||d||_\infty/2} P_d \times k(d) \times c_{i,j} - \sum_{d'=||d||_\infty/2+1}^{||d||_\infty} \underbrace{P_{d'}}_{<P_{||d||_\infty - d'}} \times k(d') \times c_{i,j} \\
&> \sum_{d=0}^{||d||_\infty/2} P_d \times c_{i,j} \times (k(d) - k(||d||_\infty - d)) \\
&> 0
\end{aligned}
$$

□

B.1.3 Proofs of Propositions 3 and 4 (monotonicity)

The Monotonicity properties 3 and 3 do not hold in general. As an illustration, fig. 7 shows the distance to the median as a function of depth for every rankings in sample generated by Mallows or Placket-Luce distributions.



(a) (b) (c) (d)

Figure 7: Each permutation in $\mathfrak{S}_6$ is a point displaying its depth (X-axis) and distance to the median (Y-axis). The ranking models are (a) Mallows model with $\phi = \mathrm{e}^{-0.375}$, (b) Mallows model with $\phi = \mathrm{e}^{-0.625}$, (c) Plackett-Luce with $\boldsymbol{w}_2 = (\mathrm{e}^n, ..., \mathrm{e}^1)$, (b) Plackett-Luce with $\boldsymbol{w}_2 = (n, ..., 1)$.

However, we derive two conditions making these monotonicity properties to hold, restricting ourselves to the case where the distance d used is Kendall τ, and also to distributions P that are strictly stochastically transitive (SST) to ensure uniqueness of the *central* ranking σ^* (see Clémençon et al. (2017)).

First, let us recall the local monotonicity property and our local monotonicity proposition:

Property 3. (Local monotonicity relative to deepest ranking) *Assume that the deepest ranking σ^* is unique. The quantity $D_P(\sigma)$ decreases as $d(\sigma^*, \sigma)$ locally increases, i.e. for any π such that $d(\sigma^*, \sigma\pi) = d(\sigma^*, \sigma)+1$, then we have $D_P(\sigma) > D_P(\sigma\pi)$.*

Proposition 3. (Local monotonicity) *If the distribution P is SST, then the Kendall τ ranking depth satisfies Property 3.*

The first part of this proposition follows immediately form Proposition 10: as we move further from the median (as measured by the Kendall τ distance) swapping adjacent ranks, the depth is strictly decreasing.

Now, we derive a second, stronger local monotonicity property. The following propostion explicit the conditions under which it is satisfied.

Proposition 14. *For a generic SST distribution P, if two rankings σ and σ' with $d = d(\sigma^*, \sigma)$ and $d' = d(\sigma^*, \sigma')$*

satisfies the following:

$$2\left[\sum_{(i,j)|p_{i,j}<1/2,\ \sigma\ correct,\ \sigma'\ incorrect} p_{i,j} - \sum_{(i,j)|p_{i,j}<1/2,\ \sigma\ incorrect,\ \sigma'\ correct} p_{i,j}\right] - (d'-d) \leq 0$$

Then the following property holds:

$$d = d(\sigma^*, \sigma) < d' = d(\sigma^*, \sigma') \Longrightarrow D_P(\sigma) \geq D_P(\sigma'), \tag{10}$$

where d is Kendall τ, σ^* is Kemeny's median and "σ correct on (i,j)" means that σ and σ^* order the pair (i,j) the same way.

The proof of this proposition can be directly derived from the proof of Proposition 4 (eq. 11)

Second, we recall the global monotonicity property and our proposition:

Property 4. (Global monotonicity) *Assume that the deepest ranking σ^* is unique. The quantity $D_P(\sigma)$ decreases as $d(\sigma^*, \sigma)$ globally increases, i.e. $d(\sigma^*, \sigma') > d(\sigma^*, \sigma) \Rightarrow D_P(\sigma') < D_P(\sigma)$.*

Proposition 4. (Global monotonicity) *If the distribution P is SST and $||d_\tau||_\infty = \binom{n}{2} < h/s$ with $h = \min_{i,j} |p_{i,j} - 1/2|$ and $s = \max_{(i,j)\neq(k,l)} |p_{i,j} - p_{k,l}|$, then the Kendall τ ranking depth satisfies Property 4.*

Proof. P is SST so $\forall\ (i,j,l), p_{i,j} > 1/2$ and $p_{j,l} > 1/2 \Longrightarrow p_{i,l} > 1/2$. WLOG, let us suppose that $\forall i < j, p_{i,j} < 1/2$. As σ^* is the unique Kemeny's median, we have $\sigma^*(n) < \sigma^*(n-1) < ... < \sigma^*(1)$ (i.e $n \succ n-1 \succ ... \succ 1$).

Let (σ, σ') be two rankings such that $d = d(\sigma^*, \sigma) < d(\sigma^*, \sigma') = d'$. Let us write $k := \#\{(i,j) | \mathbb{I}\left((\sigma^*(i) - \sigma^*(j))(\sigma(i) - \sigma(j))\right) > 0 \times \mathbb{I}\left((\sigma^*(i) - \sigma^*(j))(\sigma'(i) - \sigma'(j))\right) < 0\}$, which means that there are k pairs (i,j) on which σ agrees with σ^* (i.e. σ is "correct" on (i,j)) but σ' disagrees with σ^* (i.e. σ' is "incorrect" on (i,j)). We define k' similarly by interchanging the roles of σ and σ'.

Our goal is then to find a condition on the distribution of rankings P such that:

$$\max_{\sigma,\sigma'} L_P(\sigma) - L_P(\sigma') < 0, \qquad \text{with } k > k'$$

First, let us study the range of possible values for k. Let us divide the $n(n-1)/2$ pairs $i < j$ following:

1)	σ agrees with σ^* and σ' disagrees with σ^*	$\rightarrow k$ pairs
2)	σ agrees with σ^* and σ' agrees with σ^*	$\rightarrow a$ pairs
3)	σ disagrees with σ^* and σ' agrees with σ^*	$\rightarrow k'$ pairs
4)	σ disagrees with σ^* and σ' disagrees with σ^*	$\rightarrow b$ pairs

We then have

$$\begin{cases} k' + b = d \\ k + b = d' \\ k + a + k' + b = n(n-1)/2 \end{cases} \text{ so } \begin{cases} k' = k + d - d' \\ b = d' - k \\ a = n(n-1)/2 - k - d \end{cases}$$

Finally, since we have $0 \leq k, k', a, b \leq n(n-1)/2$, we end up having the following relevant conditions on k:

$$d' - d \leq k \leq d'$$

Now, let us write $p^{(m)}$ the m-th hightest element of the vector $(p_{i,j})_{i<j}$ of size $n(n-1)/2$, so that $1/2 > p^{(1)} > p^{(2)} > ... > p^{(n(n-1)/2)}$. Then, we have

$$
\begin{aligned}
\max_{\sigma,\sigma'} L_P(\sigma) - L_P(\sigma') &= \max_{\sigma,\sigma'} \sum_{i<j} p_{i,j} \left[\mathbb{I}\{\sigma(i)-\sigma(j)>0\} - \mathbb{I}\{\sigma'(i)-\sigma'(j)>0\}\right] + \\
&\qquad (1-p_{i,j}) \left[\mathbb{I}\{\sigma(i)-\sigma(j)<0\} - \mathbb{I}\{\sigma'(i)-\sigma'(j)<0\}\right] \\
&= \max_{\sigma,\sigma'} \sum_{i<j} (2p_{i,j}-1) \left[\mathbb{I}\{\sigma(i)-\sigma(j)>0\} + \mathbb{I}\{\sigma'(i)-\sigma'(j)<0\} - 1\right] \\
&= \max_{\sigma,\sigma'} \sum_{i<j;\sigma \text{corr.},\sigma' \text{ incorr.}} (2p_{i,j}-1) - \sum_{i<j;\sigma \text{ incorr.},\sigma' \text{ corr.}} (2p_{i,j}-1) \\
&\leq \max_{\sigma,\sigma'} 2 \left[\sum_{i<j;\sigma \text{ corr.},\sigma' \text{ incorr.}} p_{i,j} - \sum_{i<j;\sigma \text{ incorr.},\sigma' \text{ corr.}} p_{i,j}\right] - (k-k') \\
&\text{with } k' = k-(d'-d) \qquad (11) \\
&\leq 2 \left[\underbrace{p^{(1)}+\ldots+p^{(k)}}_{k \text{ elements}} - \underbrace{p^{(n(n-1)/2-k'-1)} - \ldots - p^{(n(n-1)/2)}}_{k'=k-(d'-d) \text{ elements}}\right] - (d'-d) \\
&\leq 2 \left[(p^{(1)} - p^{(n(n-1)/2-k'-1)}) + \ldots + (p^{(k')} - p^{(n(n-1)/2)}) + \right. \\
&\qquad \left. p^{(k'+1)} + \ldots + p^{(k)}\right] - (d'-d) \\
&\leq 2\left[(k' \times s + (1/2-h)(d'-d)\right] - (d'-d) \\
&\leq 2(d \times s - h) \\
&\leq 0
\end{aligned}
$$

□

Note also that if P is non-SST, then the global monotonicity property never holds, which can be easily proven by taking a counter-example and following the same proof structure.

B.2 Proof Proposition 7 (learning rate bounds)

Here we prove the finite sample results stated in the proposition below.

Proposition 7. *The following assertions hold true.*

(i) For any $\delta \in (0,1)$, we have with probability at least $1-\delta$: $\forall N \geq 1$,

$$\sup_{\sigma \in \mathfrak{S}_n} |\widehat{D}_N(\sigma) - D_P(\sigma)| \leq ||d||_\infty \sqrt{\frac{\log(2\, n!/\delta)}{2N}}.$$

(ii) For any $\delta \in (0,1)$ and $h>0$, we have with probability at least $1-\delta$: $\forall N \geq 1$,

$$\sup_{u \geq 0} |\widetilde{S}_N(u) - \widetilde{S}_P(u)| \leq \sqrt{\frac{\log(4/\delta)}{2N}} + ||d||_\infty \sqrt{\frac{\log(4n!/\delta)}{2N}}.$$

Proof. Hoeffding inequality combined with the union bound yields: $\forall t>0$,

$$\mathbb{P}\left\{\sup_{\sigma \in \mathfrak{S}_n} \left|\widehat{D}_N(\sigma) - D_P(\sigma)\right| > t\right\} \leq \sum_{\sigma \in \mathfrak{S}_n} \mathbb{P}\left\{\frac{1}{N}\left|\sum_{i=1}^N \{d(\Sigma_i,\sigma) - \mathbb{E}_P[d(\Sigma,\sigma)]\}\right| > t\right\} \leq 2n! \exp\left(-\frac{N2t^2}{||d||_\infty^2}\right),$$

which establishes assertion (i).

Turning to the proof of assertion (ii), we introduce

$$\bar{S}_P(u) = \mathbb{P}_\Sigma\{\widehat{D}_{\lfloor N/2 \rfloor}(\Sigma) \geq u\}, \quad u \geq 0.$$

By triangular inequality, we have with probability one:

$$\sup_{u\geq 0}\left|(K_h * \widehat{S}_N)(u) - (K_h * S_P)(u)\right| \leq \sup_{u\geq 0}\left|(K_h * \widehat{S}_N)(u) - (K_h * \bar{S}_P)(u)\right| + \sup_{u\geq 0}\left|(K_h * S_P)(u) - (K_h * \bar{S}_P)(u)\right|. \quad (12)$$

Observe that we almost-surely have:

$$\sup_{u\geq 0}\left|(K_h * \widehat{S}_N)(u) - (K_h * \bar{S}_P)(u)\right| \leq \sup_{u\geq 0}\left|\widehat{S}_N(u) - \bar{S}_P(u)\right|.$$

By virtue of Dvoretsky-Kiefer-Wolfovitz inequality, we have, for all $t \geq 0$,

$$\mathbb{P}\left\{\sup_{u\geq 0}\left|\widehat{S}_N(u) - \bar{S}_P(u)\right| \geq t\right\} = \mathbb{E}\left[\mathbb{P}\left\{\sup_{u\geq 0}\left|\widehat{S}_N(u) - \bar{S}_P(u)\right| \geq t \mid \Sigma_1, \ldots, \Sigma_{\lfloor N/2 \rfloor}\right\}\right] \leq 2\exp(-2nt^2). \quad (13)$$

Let $s > 0$, we introduce the event, independent from Σ,

$$\mathcal{E}_{N,s} = \left\{\sup_{\sigma\in\mathfrak{S}_n}\left|\widehat{D}_{\lfloor N/2 \rfloor}(\sigma) - D_P(\sigma)\right| \leq s\right\}.$$

We almost-surely have: $\forall u \geq 0$,

$$\bar{S}_P(u) = \mathbb{P}_\Sigma\{D_P(\Sigma) \geq u + D_P(\Sigma) - \widehat{D}_{\lfloor N/2 \rfloor}(\Sigma)\}.$$

Consequently, on the event $\mathcal{E}_{N,s}$, it holds that: $\forall u \geq 0$,

$$(K_h * S_P)(u+s) - (K_h * S_P)(u) \leq (K_h * \bar{S}_P)(u) - (K_h * \widehat{S}_N)(u) \leq (K_h * S_P)(u) - (K_h * S_P)(u-s),$$

as well as

$$\sup_{u\geq 0}\left|(K_h * S_P)(u) - (K_h * \bar{S}_P)(u)\right| \leq ||K'||_\infty (s/h), \quad (14)$$

since the mapping $K_h * S_P$ being differentiable, with derivative bounded by $||K'||_\infty/h$ in absolute value. Hence, using the union bound, combining (12) with assertion (i) and (13)-(14), we get that for all $\delta \in (0,1)$, with probability larger than $1-\delta$:

$$\sup_{u\geq 0}\left|(K_h * \widehat{S}_N)(u) - (K_h * S_P)(u)\right| \leq \left(\sqrt{\log(4/\delta)} + ||d||_\infty \sqrt{\log(4n!/\delta)}\right)/\sqrt{2N}.$$

This proves assertion (ii). □

B.3 Proofs of Propositions 5, 6 and 15 (results for Kendall τ - Mallows model)

Proposition 5. *We have:* $\forall\, \sigma \in \mathfrak{S}_n$, $D_P(\sigma) = \binom{n}{2} - \sum_{i<j} p_{i,j}\mathbb{I}\{\sigma(i) > \sigma(j)\} - \sum_{i<j}(1-p_{i,j})\mathbb{I}\{\sigma(i) < \sigma(j)\}$.

Proof. The proof is a simple computation, recalling that $\forall i \neq j, p_{i,j} = \mathbb{P}(\Sigma(i) < \Sigma(j))$. Then, $D_P(\sigma) = ||d||_\infty - \mathbb{E}_P(d_\tau(\Sigma,\sigma)) = \binom{n}{2} - \sum_{i<j}\mathbb{P}\left((\Sigma(i)-\Sigma(j))(\sigma(i)-\sigma(j)) < 0\right) = \binom{n}{2} - \sum_{i<j} p_{i,j}\mathbb{I}(\sigma(i) > \sigma(j)) - \sum_{i<j}(1-p_{i,j})\mathbb{I}(\sigma(i) < \sigma(j))$ by simple conditioning. □

Proposition 6. *Suppose that the ranking distribution P is stochastichally transitive. The following assertions hold true.*

(i) The largest ranking depth value is $D_P^* = \sum_{i<j}\left\{\frac{1}{2} + \left|p_{i,j} - \frac{1}{2}\right|\right\}$. The deepest rankings relative to P and d_τ are the permutations $\sigma \in \mathfrak{S}_n$ such that: $\forall i < j$ s.t. $p_{i,j} \neq 1/2$, $(\sigma(j) - \sigma(i)) \cdot (p_{i,j} - 1/2) > 0$.

(ii) *The smallest ranking depth value is* $\underline{D}_P = \sum_{i<j} \left\{ \frac{1}{2} - \left| p_{i,j} - \frac{1}{2} \right| \right\}$. *The least deep rankings relative to* P *and* d_τ *are the permutations* $\sigma \in \mathfrak{S}_n$ *such that:* $\forall i < j$ *s.t.* $p_{i,j} \neq 1/2$, $(\sigma(j) - \sigma(i)) \cdot (p_{i,j} - 1/2) < 0$.

(iii) *If, in addition,* P *is SST, then we have* $\partial \mathcal{R}_P(D_P^*) = \{\sigma^*\}$ *and* $\partial \mathcal{R}_P(\underline{D}_P) = \{\underline{\sigma}\}$, *where* $\sigma^*(i) = 1 + \sum_{j \neq i} \mathbb{I}\{p_{i,j} < 1/2\} = n - \underline{\sigma}(i)$ *for* $i \in \{1, \ldots, n\}$. *We also have* $D_P^* - D_P(\sigma) = 2 \sum_{i<j} |p_{i,j} - 1/2| + D_P(\sigma) - \underline{D}_P = 2 \sum_{i<j} |p_{i,j} - 1/2| \cdot \mathbb{I}\{(\sigma(j) - \sigma(i))(p_{i,j} - 1/2) < 0\}$.

Proof. Observing that $n(n-1)/2 = L_P(\sigma) + L_P(n - \sigma)$ for all $\sigma \in \mathfrak{S}_n$ in the Kendall τ case, the result is essentially a reformulation of Theorem 5 in Clémençon et al. (2017) in terms of ranking depth, insofar as $D_P = n(n-1)/2 - L_P$. □

Let us recall some classical results about the Mallows distribution. Taking $d = d_\tau$, the Mallows model introduced in Mallows (1957) is the unimodal distribution P_θ on $\mathfrak{S}_n$ parametrized by $\theta = (\sigma_0, \phi_0) \in \mathfrak{S}_n \times (0, 1]$: $\forall \sigma \in \mathfrak{S}_n$, $P_\theta(\sigma) = (1/Z_0) \exp(d_\tau(\sigma_0, \sigma) \log \phi_0)$, where $Z_0 = \sum_{\sigma \in \mathfrak{S}_n} \exp(d_\tau(\sigma_0, \sigma) \log \phi_0)$ is a normalization constant.

One may easily show that Z_0 is independent from σ_0 and that $Z_0 = \prod_{i=1}^{n-1} \sum_{j=0}^{i} \phi_0^j$. When $\phi_0 < 1$, the permutation σ_0 of reference is the mode of distribution P_{θ_0}, as well as its unique median relative to d_τ. Observe in addition that the smallest the parameter ϕ_0, the spikiest the distribution P_{θ_0}. In contrast, P_{θ_0} is the uniform distribution on $\mathfrak{S}_n$ when $\phi_0 = 1$.

A closed-from expression of the pairwise probabilities $p_{i,j}$ is available (see *e.g.* Theorem 2 in Busa-Fekete et al. (2014)). Setting $h(k, \phi_0) = k/(1 - \phi_0^k)$ for $k \geq 1$, one can then show the following: that the ranking depth function relative to P_θ and d_τ is given by:

Proposition 15. *If* $P = P_\theta$ *the Mallows distribution and* $d = d_\tau$ *the Kendall* τ *distance, then* $\forall \sigma \in \mathfrak{S}_n$, $D_{P_\theta}(\sigma) = \binom{n}{2} - \sum_{\sigma(i) > \sigma(j)} H(\sigma_0(j) - \sigma_0(i), \phi_0)$, *where* $H(k, \phi_0) = h(k+1, \phi_0) - h(k, \phi_0)$ *and* $H(-k, \phi_0) = 1 - H(k, \phi_0)$ *for* $k \geq 1$.

where $h(k, \phi_0) = k/(1 - \phi_0^k)$ for $k \geq 1$

Proof. Theorem 2 in Busa-Fekete et al. (2014) states that, for the Mallows model and using our notations, $\forall i \neq j, p_{i,j} = H(\sigma_0(j) - \sigma_0(i), \phi_0)$. The results follows from direct application of proposition 5 □

.

B.4 Proof of Proposition 8 (Borda estimators' robustness)

Proposition 8 refers to the robustness of the depth-trimmed-Borda compared to the classical Borda.

Proposition 8. *Let* μ *be the trimming threshold and* P *a distribution such that* $\mathbb{E}_P[D_P(\Sigma)] > \mu$. *Let* $\sigma^* = \arg\max_{\sigma \in \mathfrak{S}_n} D_P(\sigma)$ *be the deepest ranking and* $\pi = \arg\max_{\sigma | d_\tau(\sigma^*, \sigma) = \delta} D(\sigma)$ *the ranking with highest depth among those at distance* δ *from the deepest ranking* σ^*. *Then, the breakdown points for Borda and depth-trimmed-Borda on* P *are related as follows,*

$$\frac{\epsilon_\delta^B(P)}{\epsilon_\delta^{D\text{-}B}(P)} < \frac{D_P(\pi)}{\mu} < 1. \tag{2}$$

In this subsection, we will in fact proves some auxiliary results as well as a generalization of this proposition.

Let us first recall some definitions and results about the Borda estimators. Borda is an approximation to the barycentric ranking median (which is NP-hard for $n > 4$ Dwork et al. (2001)) for a sample of complete rankings drawn from a MM Fligner and Verducci (1988). Moreover, Borda is quasi-linear in time and outputs the correct median w.h.p. with a polynomial number of samples Caragiannis et al. (2013). A robust aggregation procedure for top-k rankings in very noisy settings is proposed in Collas and Irurozki (2021).

The Borda median estimator for sample X orders the items increasingly by their Borda score, defined as $B(i) = \sum_{\sigma \in X} \sigma(i)$.

We define the depth-weighted-Borda as a generalization of the classic and depth-trimmed-Borda in which there exists a weight associated with each ranking. It generalizes Borda in the following way: For each item i, the

Borda score is computed as $B(i) = \sum_{\sigma \in X} w(\sigma)\sigma(i)$. The final estimator for the median is the ranking that orders the items by their Borda score. The depth-weighted-Borda is equivalent to replicating the rankings proportionally to their weight. This analysis generalizes to any weights are *increasing* function of the depths. In particular, the depth-trimmed-Borda is the case of depth-weighted-Borda in which $w(\sigma) = \mathbb{I}\{D(\sigma) > \mu\}$.

We settle here the notation for the following lines. We denote by $S_N \sim P$ a sample of rankings (of size N) and A an adversarial sample.

Definition 7. *Let P be a distribution, let us write $S_N \sim P$ a sample drawn from P of size N and $\sigma^T_{S_N}$ the median based on the estimator method T on sample S_N.*

The estimator T is said to be δ-broken (for Kendall's τ) for sample size N and distribution P if for any $S_N \sim P$ of size N, there exists an adversarial sample A such that $d_\tau(\sigma^T_{S_N}, \sigma^T_{S_N \cup A}) \geq \delta$.

The next result characterizes the carnality of a sample that breaks the Borda estimator of a sample S_N distributed according to P. This is an auxiliary result for Proposition 8.

Proposition 16. *Let $S_N \sim P$. Let A^- be the adversarial sample that δ-breaks the Borda estimator (for sample size N and distribution P) such that A^- is of minimal cardinality. Let $\bar{r}_N(i) = N^{-1}\sum_{\sigma \in S_N} \sigma(i)$ and $\bar{r}(i) = (\#A^-)^{-1}\sum_{\sigma \in A^-} \sigma(i)$ be the average ranking of item i in S_N and A^- respectively. Finally, let $\bar{R}$ be the ordered vector composed of $\frac{\bar{r}_N(j) - \bar{r}_N(i)}{\bar{r}(i) - \bar{r}(j)}$ for all (i,j) such as both the numerator and denominator are positive. Then*

$$\#A^- = \lceil N \left[\bar{R}\right]_{(\delta)} \rceil$$

where $[x]_{(\delta)}$ denotes the δ-th quantile of a vector x.

Proof. By definition, A^- δ-breaks Borda iff the following holds.

$$\begin{aligned}
&d(\sigma^B_{S_N}, \sigma^B_{S_N \cup A^-}) = \delta \\
&\quad \Leftrightarrow \delta = \#\{(i<j) : \sum_{\sigma \in S_N} \sigma(i) + \sum_{\sigma \in A^-} \sigma(i) \geq \sum_{\sigma \in S_N} \sigma(j) + \sum_{\sigma \in A^-} \sigma(j)\} \\
&\quad \Leftrightarrow \delta = \#\{(i<j) : \sum_{\sigma \in S_N} \sigma(i) - \sigma(j) \geq \sum_{\sigma \in A^-} \sigma(j) - \sigma(i)\} \\
&\quad \Leftrightarrow \delta = \#\{(i<j) : \sum_{\sigma \in S_N} \sigma(j) - \sigma(i) \leq \sum_{\sigma \in A^-} \sigma(i) - \sigma(j)\} \\
&\quad \Leftrightarrow \delta = \#\{(i,j) : 0 < \sum_{\sigma \in S_N} \sigma(j) - \sigma(i) \leq \sum_{\sigma \in A^-} \sigma(i) - \sigma(j)\}
\end{aligned} \tag{15}$$

From a statistical perspective, we can bound the cardinality of A^- as follows: let (i,j) be a pair of index belonging to the set define just above.

$$\begin{aligned}
&\sum_{\sigma \in S_N} \sigma(j) - \sigma(i) \leq \sum_{\sigma \in A^-} \sigma(i) - \sigma(j) \\
&\Leftrightarrow N\left(\bar{r}_N(j) - \bar{r}_N(i)\right) \leq \#A^-(\bar{r}(i) - \bar{r}(j)) \\
&\Rightarrow \#A^- \geq \frac{N\left(\bar{r}_N(j) - \bar{r}_N(i)\right)}{\bar{r}(i) - \bar{r}(j)},
\end{aligned} \tag{16}$$

which holds for exactly δ pairs of items (i,j). We conclude the proof by recalling that A^- is of minimal cardinality. □

The next auxiliary result shows that provided certain conditions, if a sample breaks the depth-weighted-Borda then it breaks Borda.

Proposition 17. *Let $S_N \sim P$. Let A^- (resp. A^-_w) be the adversarial sample that δ-breaks the Borda (resp. depth-weighted Borda) estimator (for sample size N and distribution P) such that A^- (resp. A^-_w) is of minimal cardinality. Let $\bar{r}_N(i) = N^{-1}\sum_{\sigma \in S_N} \sigma(i)$ and $\bar{r}_w(i) = (\#A^-_w)^{-1}\sum_{\sigma \in A^-_w} \sigma(i)$ be the average ranking of item i in S_N and A^-_w respectively.*

Let $\pi_w = \arg\max_{\sigma \in A_w^-} w(\sigma)$ *and* $\mu = w(\pi_w)$ *the threshold of maximum depth for adversarial rankings.*

Finally, suppose that $\hat{P}_N$ *and* w *satisfy:* $\mathbb{E}_{\hat{P}_N}(w(\Sigma)) > w(\pi_w) = \mu$ *and* $\forall\ (i,j)$ *s.t.* $\mathbb{E}_{\hat{P}_N}(\Sigma(j) - \Sigma(i)) > 0$, $\mathbb{E}_{\hat{P}_N}(w(\Sigma)(\Sigma(j) - \Sigma(i)) \geq \mathbb{E}_{\hat{P}_N}(w(\Sigma))\mathbb{E}_{\hat{P}_N}(\Sigma(j) - \Sigma(i))$ *(these two assumptions enforce the use of a weight function that is in accordance with* $\hat{P}_N$*)*

Then, the cardinality of A^- *and* A_w^- *are related as follows:*

$$\#A_w^- \geq \frac{N^{-1}\sum_{\sigma \in S_N} w(\sigma)}{\mu} \#A^-.$$

Proof. Since A_w^- δ-breaks the depth-weighted-Borda, we can follow the same proof outline as for proposition 16 and bound the cardinality $\#A_w^{-1}$ as follows,

$$\begin{aligned} &\sum_{\sigma \in S_N} w(\sigma)(\sigma(j) - \sigma(i)) \leq \sum_{\sigma \in A_w^-} w(\sigma)(\sigma(i) - \sigma(j)) \\ &\Rightarrow N \times\ N^{-1} \sum_{\sigma \in S_N} w(\sigma)(\sigma(j) - \sigma(i)) \leq \#A_w^- w(\pi)(\bar{r}_w(i) - \bar{r}_w(j)) \\ &\Rightarrow \#A_w^- \geq \frac{N\,(\bar{r}_n(j) - \bar{r}_N(i))}{\bar{r}_w(i) - \bar{r}_w(j)} \frac{N^{-1}\sum_{\sigma \in S_N} w(\sigma)}{\mu} \end{aligned} \tag{17}$$

Since $\frac{N^{-1}\sum_{\sigma \in S_N} w(\sigma)}{\mu}$ is independent of i, j and A_w^- also δ-breaks the Borda estimator, we can conclude:

$$\#A_w^- \geq \#A^- \frac{N^{-1}\sum_{\sigma \in S_N} w(\sigma)}{\mu}. \tag{18}$$

□

We are finally ready to prove a generalization of our proposition 8 stated in the main paper. Let us first define our notion of *δ-breakdown point*, which extends the classical concept.

Definition 8. *Let* P *be a distribution. The* δ*-breakdown point for an estimator* T *with respect to distribution* P *is defined as the smallest cardinality of an adversarial sample that* δ*-breaks* T *in the limit when* $N \to \infty$ *for distribution* P*.*

More specifically, $\epsilon_\delta^T(P) = \min \#A$ *s.t.* $\lim_{N\to\infty} d_\tau(\sigma^T_{S_N}, \sigma^T_{S_N \cup A}) = \delta$

In the following proposition, we write $\epsilon_\delta^B(P)$ (resp. $\epsilon_\delta^{DW-B}(P)$) the δ-breakdown point for the Borda (resp. depth-weighted Borda) estimator with respect to distribution P.

Proposition 18 (breakdown points ratio)**.** *Let* P *be a distribution such that* $\mathbb{E}_P[w(\Sigma)] > w(\pi)$*, where* $\pi = \arg\max_{\sigma \mid d_\tau(\sigma^*, \sigma) = \delta} w(\sigma)$ *and* $\sigma^* = \arg\max_{\sigma \in \mathfrak{S}_n} D_P(\sigma)$*. Let* P *and* w *satisfy:* $\forall (i,j)$ *s.t.* $\mathbb{E}_P(\Sigma(j) - \Sigma(i)) > 0$, $\mathbb{E}_P(w(\Sigma)(\Sigma(j) - \Sigma(i))) \geq \mathbb{E}(w(\Sigma))\mathbb{E}(\Sigma(j) - \Sigma(i))$*. Then,*

$$\lim_{N\to\infty} \frac{\epsilon_\delta^B(P)}{\epsilon_\delta^{DW-B}(P)} < \frac{w(\pi)}{\mathbb{E}_P[w(\Sigma)]} < 1. \tag{19}$$

Proof. We start by noting that for S_N to be δ-broken then the adversarial sample has to be at least at distance δ regardless the distribution for the weights. Then, we denote $z = \mathbb{E}_P[w(\Sigma)]/w(\pi) = \lim_{N\to\infty} N^{-1}\sum_{\sigma \in S_N} w(\sigma)/w(\pi)$ (by the law of large numbers) and take Proposition 17 to write the limiting ratio of the breakdown points when the number of samples tends to infinity as follows.

$$\lim_{N\to\infty} \frac{\epsilon_\delta^B(P)}{\epsilon_\delta^{DW-B}(P)} = \lim_{N\to\infty} \frac{\frac{\#A^-}{\#A^- + N}}{\frac{\#A_w^-}{\#A_w^- + N}} < \lim_{N\to\infty} \frac{\frac{\#A^-}{\#A^- + N}}{\frac{\#A^- \cdot z}{\#A^- \cdot z + N}} < \frac{1}{z} = \frac{w(\pi)}{\mathbb{E}_P[w(\Sigma)]} < 1 \tag{20}$$

□

This is the main result related to the robustness of the Borda median estimator. It shows that the breakdown point of Borda is smaller than the breakdown point for the depth-trimmed-Borda provided certain conditions. We denote by μ the threshold of the depth-trimmed-Borda.

Then, our proposition 8 is straightforward when we choose the weight function w so that $w(\sigma) = \mathbb{I}(D_P(\sigma) \geq \mu)$ in Proposition 18.

C Further results

C.1 Ranking quantile function

In Proposition 7, rate bounds for the deviation between empirical and theoretical versions of the depth function (respectively, of the smoothed depth survivor function) have been stated. We here give some indications for obtaining similar results for the ranking quantile function.

As the considered distribution is discrete, finite and real-valued (because depth is real-valued), the results of Ma et al. (2011) can be directly applied. Using their notations, we have:

Σ is a random variable of distribution P on $\mathfrak{S}_n$ that takes distinct values $\sigma_1, \ldots, \sigma_m$ with respective propabilities $\rho_1, \ldots, \rho_m$. Each element $\sigma_i \in \{\sigma_1, ..., \sigma_m\}$ has a depth δ_i, which leads us to write $D_P(\Sigma)$ the random variable associated to the depth.

Now, let us reorder the indices and write the distinct depth values $\delta_1 < ... < \delta_d$ with respective probabilities of occurrence $p_1, \ldots, p_d$, where $\forall\ i \in [\![1, d]\!], p_i = \sum_{1 \leq j \leq m} \rho_j \mathbb{1}(\delta_j = \delta_i)$. Let us define the *mid-function* $F_{mid}(x) = \mathbb{P}(D_P(\Sigma) \leq x) - 1/2\mathbb{P}(D_P(\Sigma) = x)$ and the ranking quantile function based on *mid-functions*:

$$Q(\alpha) = F^{-1}_{mid}(\alpha) = \begin{cases} \delta_1 & \text{if } \alpha \leq p_1/2 \\ \lambda\delta_k + (1-\lambda)\delta_{k+1} & \text{if } \alpha = \lambda\pi_k + (1-\lambda)\pi_{k+1} \\ & \quad \text{for any } \lambda \in [0,1] \text{ and } 1 \leq k \leq d-1 \\ \delta_d & \text{if } \alpha \geq \pi_d \end{cases}, \tag{21}$$

where $\forall k \in [\![2, d]\!], \pi_k = \sum_{i=1}^{k-1} p_i + p_k/2 = F_{mid}(\delta_k)$.

Then, the following results hold:

$$\begin{aligned}
&1)\ \hat{Q}_N(\alpha) \xrightarrow{\mathbb{P}} \delta_1 \text{ if } \alpha < p_1/2 \\
&2)\ \hat{Q}_N(\alpha) \xrightarrow{\mathbb{P}} \delta_d \text{ if } \alpha > \pi_d \\
&3)\ \sqrt{N}(\hat{Q}_N(\alpha) - (\lambda\delta_{k+1} + (1-\lambda)\delta_{k+2})) \xrightarrow{\mathbb{P}} \mathcal{N}(0, sd(\alpha, \lambda, p_{k+1}, p_{k+2})) \\
&\qquad \text{if } \alpha = \lambda\pi_{k+1} + (1-\lambda)\pi_{k+2} \text{ for } 0 < \lambda < 1 \text{ and } 0 \leq k \leq d-2 \\
&4)\ \sqrt{N}(\hat{Q}_N(\alpha) - \delta_{k+1}) f(\hat{Q}_N(\alpha), \delta_{k+1}) \xrightarrow{\mathbb{P}} \mathcal{N}(0, \alpha(1-\alpha) - p_{k+1}/4) \\
&\qquad \text{if } \alpha = \pi_{k+1} \text{ for } 0 \leq k \leq d-1,
\end{aligned}$$

where $sd(\alpha, \lambda, p_{k+1}, p_{k+2}) = \alpha(1-\alpha) - (1-(\lambda-1)^2)p_{k+1}/4 - (1-\lambda^2)p_{k+2}/4$ and $f(\hat{Q}_N(\alpha), \delta_{k+1}) = 1/2(p_{k+1} + p_{k+2})/(\delta_{k+2} - \delta_{k+1})$ if $\hat{Q}_N(\alpha) > \delta_{k+1}$ and $1/2(p_{k+1} + p_k)/(\delta_{k+1} - \delta_k)$ else.

These results provide us with asymptotic guarantees about the ranking quantile function based on mid-functions, as defined in eq. 21. However, non-asymptotic bounds as well as similar results for the depth regions should be investigated further and are left for future work, like the discrepancy between empirical and theoretical ranking depth regions, which can be measured by *e.g.* the cardinality of their symmetric difference.

C.2 Pairwise comparisons as an alternative statistical framework

Since the computation of Kendall τ distance involves pairwise comparisons only, one could compute empirical versions of the risk functional L in a statistical framework stipulating that the observations are less complete

than $\{\Sigma_1, \ldots, \Sigma_N\}$ and formed by i.i.d. pairs:

$$(\mathbf{e}_k, \epsilon_k),\ k = 1, \ldots, N,$$

where the $\mathbf{e}_k = (\imath_k, \jmath_k)$'s are independent from the Σ_k's and drawn from an unknown distribution ν on the set $\mathcal{E}_n$ such that $\nu(e) > 0$ for all $e \in \mathcal{E}_n$ and $\epsilon_k = sgn(\Sigma_k(\jmath_k) - \Sigma_k(\imath_k))$ with $\mathbf{e}_k = (\imath_k, \jmath_k)$ for $1 \leq k \leq N$. Based on these observations, an estimate of the risk of any median candidate $\sigma \in \mathfrak{S}_n$ is given by:

$$\widetilde{L}_N(\sigma) = \sum_{i<j} \frac{1}{N_{i,j}} \sum_{k=1}^{N} \mathbb{I}\{\mathbf{e}_k = (i,j),\ \epsilon_k(\sigma(j) - \sigma(i)) < 0\}, \tag{22}$$

where $N_{i,j} = \sum_{k=1}^{N} \mathbb{I}\{\mathbf{e}_k = (i,j)\}$.

D Additional experiments

Here we display additional numerical results, completing those presented in the main text. First, in Section **??**, we analyze the sensitivity of the proposed depth notion to a difference between distributions and its subsequent ability to provide formal inference. Second, in Section **??**, we detail a further application to real data. Please note, that, as this is the case in the main text, in all visualizations the data depth is re-scaled to $[0, 1]$ interval by division by maximal possible distance for given n.

D.1 Trimming strategy

Now we have characterized under which conditions the different properties of Property **??** hold, we explore how to use them in practice. For example, using Kendall τ distance and samples drawn from a Mallows distribution easily make the invariance and maximality at the center properties hold, but not necessarily the monotonicity property.

First, even though a Mallows model is SST, its empirical distribution counterpart may not be. Second, the adjacent condition for local monotonicity indicates that under such a model,the monotonicity property hold for the median and any of its adjacent ranking. Moreover, the second local condition is more likely to be satisfied for rankings close to (in terms of Kendall τ distance) the median.

These two observations and the fact that the depth of a ranking σ represents its centrality within the dataset make a *trimming strategy* highly relevant for consensus ranking experiments. The intuition behind this strategy is that least deep points corresponds to *outliers* for the dataset: removing them step by step should make the dataset less noisy, so that the depth of the remaining points get more and more accurate. To the extreme, when successively trimming rankings in the dataset until there is only one ranking left should leave us with an accurate median for the dataset. However, since the depth function satisfies useful properties when the underlying distribution is SST, a sufficient trimming strategy would stop there. The algorithm corresponding to this strategy is defined in Algorithm 1 of the main paper, recalled here.

Algorithm 2: Ranking Depth Trimming

Input : Ranking dataset $\mathcal{D}_N = \{\Sigma_1, ...\Sigma_N\}$ and distribution $\widehat{P}_N = (1/N)\sum_{i=1}^{N} \delta_{\Sigma_i}$.
Output : Dataset $\mathcal{D} \subset \mathcal{D}_N$ of size $N_{\mathcal{D}} \leq N$ and (S)ST ranking distribution $\widehat{P}_{\mathcal{D}} = (1/N_{\mathcal{D}})\sum_{\sigma \in \mathcal{D}} \delta_\sigma$
- Initialize: $\mathcal{D} = \mathcal{D}_N$;
while $\hat{P}_{\mathcal{D}}$ *is not (S)ST* **do**
  - Determine the least deep rankings in $\mathcal{D}$: $\mathcal{O}_{\mathcal{D}} := \arg\min_{\sigma \in \mathcal{D}} D_{\hat{P}_N}(\sigma)$;
  - Update the ranking dataset $\mathcal{D} \setminus \mathcal{O}_{\mathcal{D}} \to \mathcal{D}$

Fig. 8 illustrates the trimming strategy for a Mallows model generated with $n = 8$ items, $\phi_0 = 0.985$, and $N = 1000$ samples. We can see that trimming indeed remove cycles from the empirical dataset and thus make the empirical distribution SST, and that during trimming, the deepest rankings (saved as the candidate medians) get closer (in Kendall τ distance) to the real median used for generating the samples.

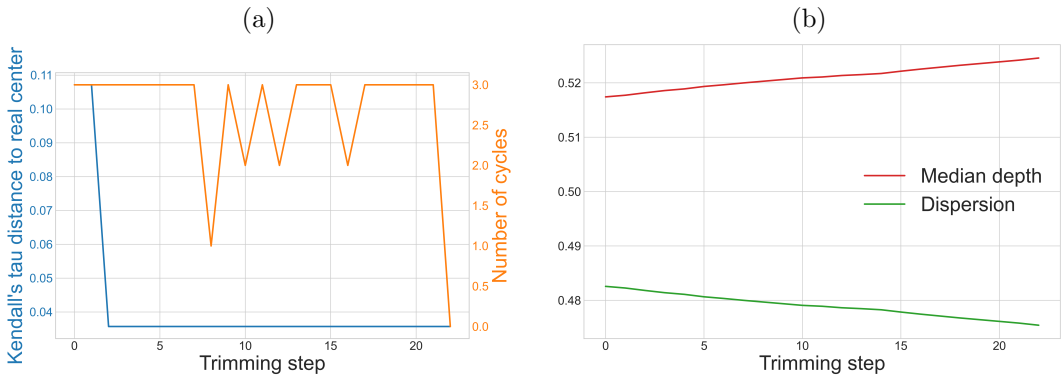


Figure 8: Trimming strategy: evolution of candidate median (deepest ranking) normalized distance to real median and number of cycles through trimming (a); evolution of median depth and sample dispersion through trimming (b).

The depth function thus provide an alternative and relevant way to compute the median of a dataset: by trimming until getting a SST distribution first, we ensure that the depth function has desirable properties and thus that the median we obtain in practice gets very close to the true median.

D.2 Visual analysis

With data depth being a nonparametric tool not exploiting *a priori* information about the distribution, we focus on easy-to-manipulate Mallows model using Kendall τ distance; we refer to the main text for the formal definition and parameters' notation, see Example 1.

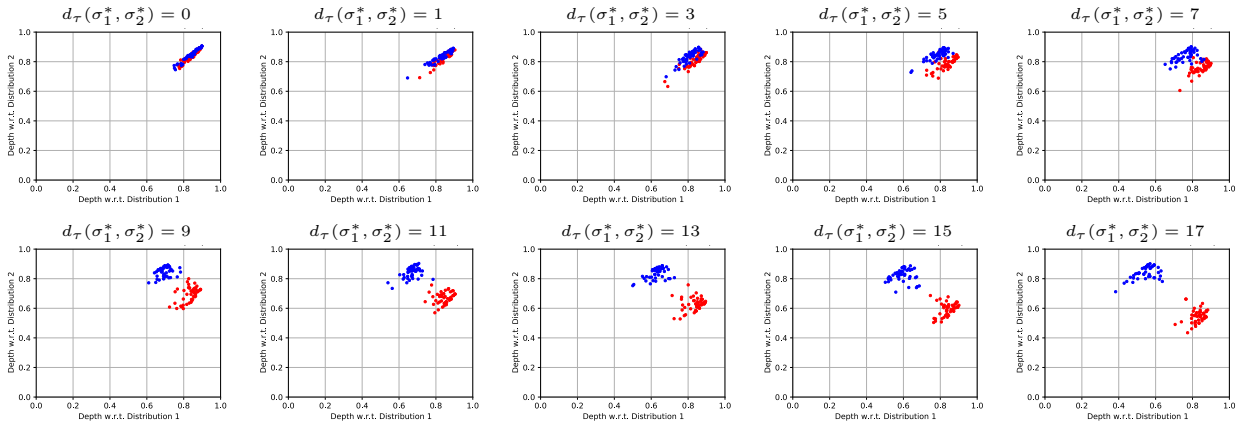


Figure 9: *DD*-plots for pairs of distributions stemming from different instances of the *location-shift model*. The two distributions contain 50 observations each, drawn from two Mallows models using Kendall τ distance with parameter $\phi_1 = \phi_2 = \mathbf{e}^{-1}$. Difference between locations is indicated in each individual plot.

First, we consider a *location-shift model*: a sequence of pairs of distributions with parameter pairs (σ_1^*, ϕ_1) and (σ_2^*, ϕ_2). Setting $\phi_1 = \phi_2 = \mathbf{e}^{-1}$, we vary σ_2^* so that $d_\tau(\sigma_1^*, \sigma_2^*) \in \{0, 1, 3, 5, 7, 9, 11, 15, 17\}$. Figure 9, which contains visualization for each pair of parameters for 50 observations from each distribution, illustrates gradual capturing by the suggested visualization of the increasing location shift between the two laws.

Bearing in mind the same idea, using the same distributional settings, we now provide a formal statistical inference by *homogeneity testing*. More precisely, for each pair of distributions, taking one of them for a reference, we perform the testing procedure 100 times and indicate average p-values in Figure 10 (where we stick to the same test setting as in the main text, drawing 500 observations for the reference distribution and using the Wilcoxon rank-sum statistic). As expected, when there is no parameter difference, the null hypothesis of the distribution's

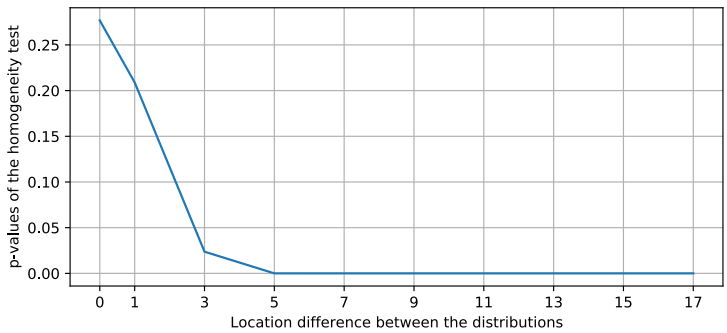


Figure 10: p-values (averaged over 100 random repetitions) for the test of homogeneity for a pair of Mallows distributions stemming from the *location-shift model* with location difference $d_\tau(\sigma_1^*, \sigma_2^*) \in \{0, ..., 17\}$.

equality cannot be surely rejected. When the difference in the location increases, it is captured very quickly by the testing procedure rejecting on the level ≤ 0.05 when $d_\tau(\sigma_1^*, \sigma_2^*) = 3$ only, and with even higher reliability for larger differences.

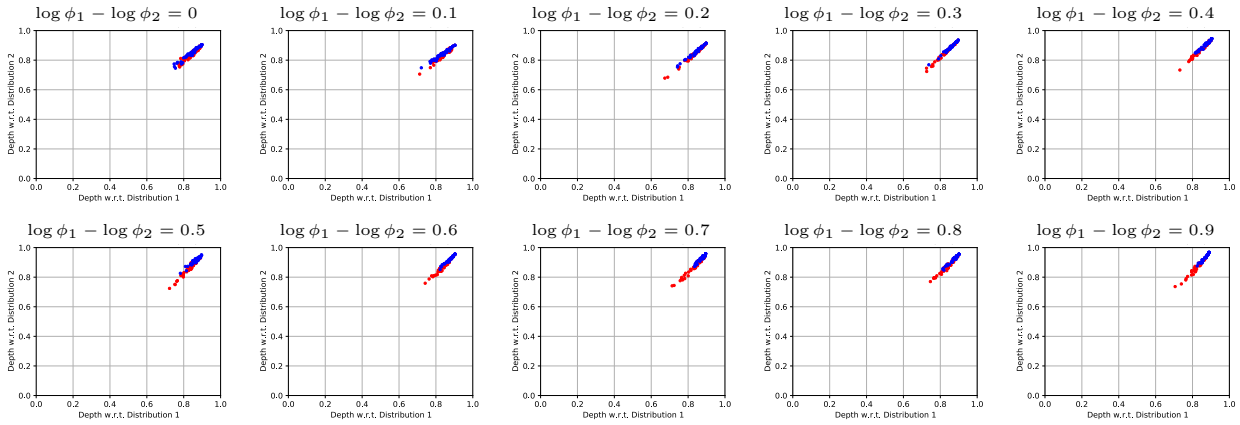


Figure 11: *DD*-plots for pairs of distributions stemming from different instances of the *scale-difference model*. The two distributions contain 50 observations each, drawn from two Mallows models (using Kendall τ distance) with the same center and with parameters $\phi_1 = \mathbf{e}^{-1}$ and $\phi_2 = \mathbf{e}^{\psi}$ where $\psi \in \{-1, -1.1, -1.2, -1.3, -1.4, -1.5, -1.6, -1.7, -1.8, -1.9\}$. Difference between scales' logarithms is indicated in each individual plot.

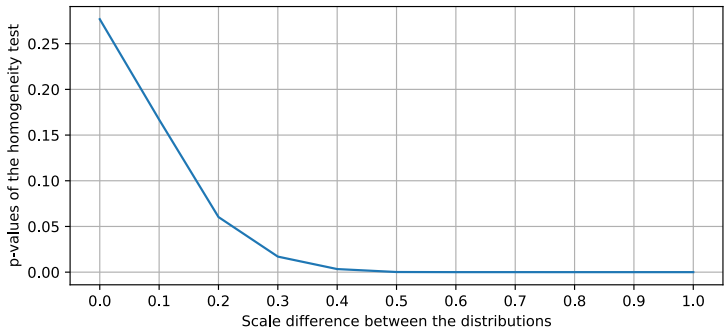


Figure 12: p-values (averaged over 100 random repetitions) for the test of homogeneity for a pair of Mallows distributions stemming from the *scale-difference model*, *i.e.* with the same center and difference in scales $\log \phi_1 - \log \phi_2 \in \{0, ..., 1\}$.

While being already challenging when having no (parametric) assumptions on the distribution, traditional rank-based homogeneity testing procedures usually assume location difference. Thus, we consider an even more disadvantageous setting, the *scale-difference model*: the location of the both distributions is the same, and those differ in dispersion only. As above, the DD-plots for 10 scale difference values (on the equidistant grid with step 0.1 on the logarithmic scale) are presented in Figure 11. With increasing scale difference, visual patterns dis-associate (less than in the previous setting though) which intrigues the formal inference.

Finally, we repeat the previously used testing procedure and indicate the average p-values in Figure 12. One observes that for difference in scale (measured on the logarithmic scale) equal to 0.3 or higher, the homogeneity testing procedure distinguishes the distributions with level ≤ 0.05 or less.

D.3 Application to real data

Let us now explore the applicability of our depth function to different tasks on real data.

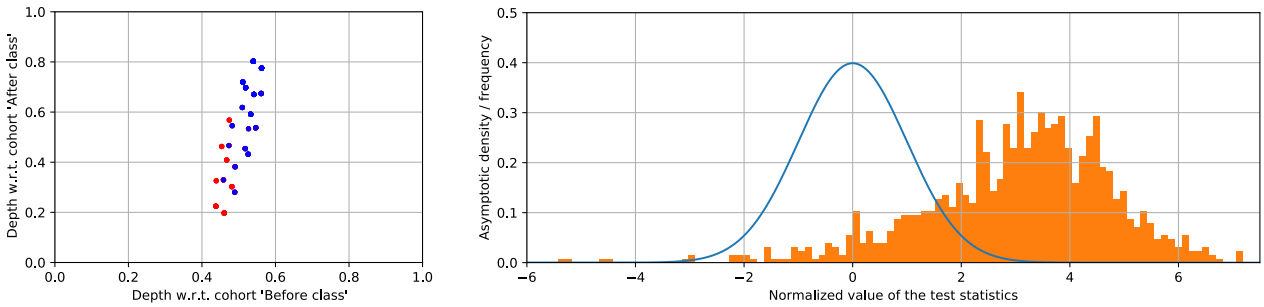


Figure 13: Left: DD-plot for the two cohorts of students, 'before class' (red) and 'after class' (blue), respectively. Right: p-values of the homogeneity test over 1000 random repetitions together with asymptotic density under H_0.

D.3.1 Student dataset

Next, we consider a real data set which consists of students' rankings before ($N_1 = 169$ students) and after ($N_2 = 179$ students) the class, with known ground truth (correct answer $= (0, 1, 2, 3)$) where $n = 4$ (refer to `https://github.com/ekhiru/students-dataset` for details about the dataset). Simple computation indicates that d_τ from the average ranking to the true one is 2 before the class and 1 after, thus suggesting that the students improved after studying. We employ the same homogeneity testing methodology as above to derive formal statistical inference. By taking 100 randomly chosen observations from the 'before class' cohort as the reference, we use 69 observations from the 'before class' cohort and 79 (also randomly chosen) from the 'after class' group. The diagnostic DD-plot of the two cohorts together with p-values over 1000 random repetitions and the asymptotic density under H_0 are indicated in Figure 13, and illustrate improvement of the students' knowledge after the class.

D.3.2 Sushi dataset

As a last application, we analyze the *Sushi* data set, which contains 5 000 rankings of 10 sushi items. We refer the reader to Kamishima (2013) (`https://www.kamishima.net/sushi/`) for the detailed description of the data set. By means of the introduced depth notion, we explore this data set from two angles. First, we provide depth based ranking of the entire data set, which can be seen as the ranking equivalent of the cumulative distribution function. The depth of each of the 5 000 observations (ordered increasingly) is indicated in Figure 14 (left). Second, in view of the ten considered items, we check the know difference between food preferences in Eastern and Western Japan. The DD-plot of these two groups (containing 3 448 and 1 552 observations each, respectively) is presented in Figure 14 (right). Since the two clouds of points substantially intersect, this rather drives to conclusion that the mentioned above difference is not connected with the choice of the sushi items used in the data set.

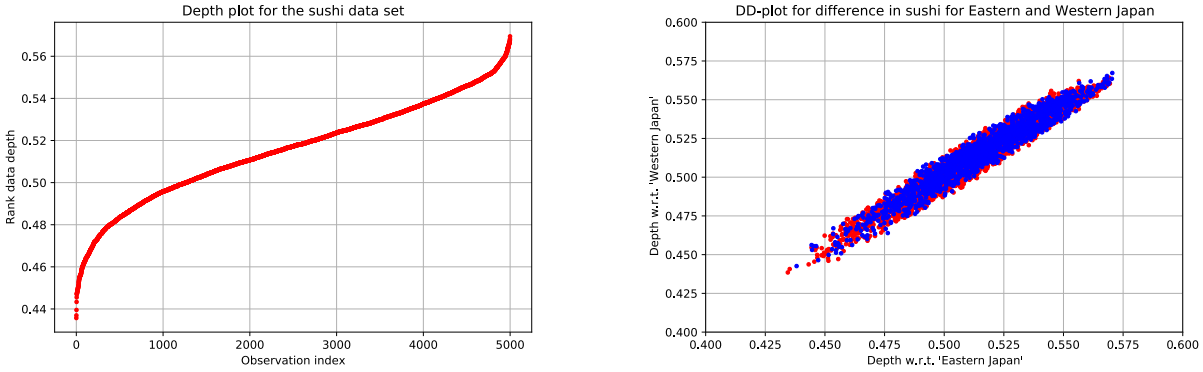


Figure 14: Exploratory statistics of the Sushi data set by Kamishima (2013) (`https://www.kamishima.net/sushi/`). Left: The depth of each observation in the set, in increasing depth order. Right: The comparative DD-plot for Eastern (red) and Western (blue) Japan.

D.3.3 Mechanical Turk Dots dataset

The Mechanical Turk Dots dataset contains 800 full rankings of 4 items. Each item corresponds to random dots presented to a user on Mechanical Turk, who is asked to rank them from those containing the least dots (first) to those containing the most dots (last). Thus, there is a ground truth ranking for this dataset. 40 sets of puzzles were placed on Mechanical Turk and were ranked by 20 users, leading to 800 rankings.

This dataset is SST and the deepest ranking corresponds to the ground truth. We thus contaminate the dataset by swapping a random proportion of 1/4 of the rankings, i.e. by taking the opposite ranking. Figure 15 (a) shows that there is no obvious difference between the swapped and clean rankings, but in figure 15 (b), we see we recovered the ground truth ranking after the trimming strategy.

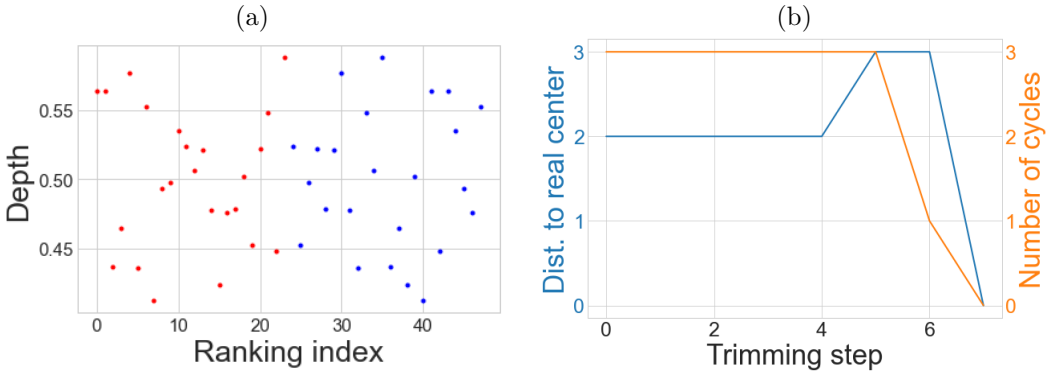


Figure 15: Depth plots before trimming with swapped (red) and clean (blue) points; evolution of candidate median (deepest ranking) distance to real median and number of cycles through trimming (c)

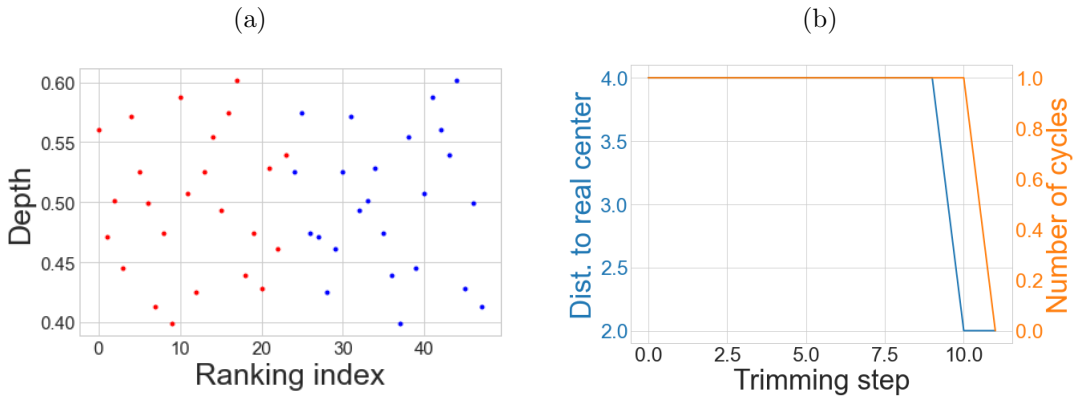


Figure 16: Depth plots before trimming with swapped (red) and clean (blue) points; evolution of candidate median (deepest ranking) distance to real median and number of cycles through trimming (c)

D.3.4 Netflix Prize dataset

We selected one of the Netflix Prize dataset contains 1814 full rankings of 4 movies (Dirty Dancing, Maid in Manhattan, Shrek and Father of the Bride). This dataset is SST and the deepest ranking corresponds to Shrek $\succ$ Father of the Bride $\succ$ Maid in Manhattan $\succ$ Dirty Dancing, considered as the real center of the dataset. We contaminated the dataset by swapping a random proportion of 11% of the rankings, i.e. by taking the opposite ranking. Figure 16 (a) shows that there is no obvious difference between the swapped and clean rankings, but in figure 16 (b), we the median computed after trimming is closer to the real center than before.