Beyond Kemeny Medians: Consensus Ranking Distributions. Definition, Properties and Statistical Learning

Stephan Clémençon[1] **Ekhine Irurozki**[1,2]

[1]LTCI, Télécom Paris, Institut Polytechnique de Paris, France
[2]University of the Basque Country (UPV/EHU), Spain

{stephan.clemencon, irurozki}@telecom-paris.fr

Abstract

In this article we develop a new method for summarizing a ranking distribution, *i.e.* a probability distribution on the symmetric group $\mathfrak{S}_n$, beyond the classical theory of consensus and Kemeny medians. Based on the notion of *local ranking median*, we introduce the concept of *consensus ranking distribution* (CRD), a sparse mixture model of Dirac masses on $\mathfrak{S}_n$, in order to approximate a ranking distribution with small distortion from a mass transportation perspective. We prove that by choosing the popular Kendall τ distance as the cost function, the optimal distortion can be bounded in terms of pairwise probabilities, paving the way for the development of efficient learning methods that do not suffer from the lack of vector space structure on $\mathfrak{S}_n$. In particular, we propose a top-down tree-structured statistical algorithm that allows for the progressive refinement of a CRD based on ranking data, from the Dirac mass at a Kemeny median at the root of the tree to the empirical ranking data distribution itself at the end of the tree's exhaustive growth. In addition to the theoretical arguments developed, the relevance of the algorithm is empirically supported by various numerical experiments.

1 Introduction

Designing machine learning algorithms tailored to preference data is key to optimizing the performance of recommendation systems and search engines. These tools, which are indispensable in the digital age, continuously use and generate ever-increasing volumes of data. The nature of the data feeding into or produced by these algorithms, which mainly consists of (partial) rankings expressing observed or predicted preferences, continues to pose a methodological challenge. Given that the number $n!$ of possible rankings explodes with the number $n \geq 1$ of instances to be ordered, it is crucial to develop dedicated dimensionality reduction methods in order to represent/analyze ranking data in an efficient manner. Regardless of the type of task being considered (supervised, unsupervised), the vast majority of machine learning algorithms based on multivariate data rely on the calculation of statistical quantities such as (pairwise) averages or linear combinations of features, effectively representing (location and dispersion) properties of the data distribution. Summarizing ranking distributions, *i.e.* probability laws on the group of permutations $\mathfrak{S}_n$, is anything but simple. Given the absence of a vector space structure on $\mathfrak{S}_n$, extending basic concepts such as mean/median or variance, ubiquitous in multivariate data analysis, to preference data poses mathematical and computational difficulties.

Although it is obviously possible, based on the prior choice of a metric on $\mathfrak{S}_n$ (most often the Kendall τ distance, *i.e.* the number of discording pairs), to define a barycentric permutation σ^* in $\mathfrak{S}_n$, *i.e.* a ranking median, and a related dispersion measure (the average distance to the median namely)

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

from a collection of rankings, its calculation can prove difficult. This problem, known as *consensus ranking* or *ranking aggregation* and referring to the pioneering work of Condorcet in social choice theory, continues to be the subject of intense interest, one may refer to *e.g.* [10], [12], [19], [20] or [31] among others. In practice, ranking medians provide a summary of a statistical population of rankings that is far too crude in many cases, and rigid statistical modeling using classical probability laws on $\mathfrak{S}_n$ (*e.g.* Mallows [12], Placket-Luce [6]) reaches its limits as soon as ranking data exhibit significant heterogeneity. To move beyond these rigid parametric models, statistical depth functions have recently been proposed [15] for ranking distributions, while [13] define depth functions within Formal Concept Analysis for the more general setting of partial orders. In this paper, we likewise build on depth functions, but extend their use to heterogeneous, multimodal distributions.

A concrete illustration of heterogeneous distribution of rankings is benchmarking of LLMs, see [5]: each task (*e.g.* translation, text generation, summarization, ...) generates a ranking of LLMs according to their performance on the given task. Different tasks rank systems differently, so the resulting distribution over rankings is naturally multimodal, and a single consensus ranking discards exactly the structure a practitioner would want surfaced. We use the TAC-08 benchmark from this line of work as a running illustration throughout the paper.

In this article, we develop a new framework for approximating/estimating a ranking distribution P, *i.e.* the law of a generic r.v. $\Sigma = (\Sigma(1), \ldots, \Sigma(n))$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and valued in $\mathfrak{S}_n$, extending the concept of Kemeny medians, *i.e.* ranking medians when the metric chosen on $\mathfrak{S}_n$ is the Kendall τ distance. Our approach is inspired by *local averaging methods*, which are widely used in multivariate data analysis and underlie many popular methods and algorithms such as K-means or histogram estimation. It is based on the concept of *local median* which, combined with a partition $\mathcal{P}$ of the symmetric group and associated class probabilities $\alpha_{\mathcal{C}} = \mathbb{P}\{\Sigma \in \mathcal{C}\}$ with $\mathcal{C} \in \mathcal{P}$, allows us to define what we call a *consensus ranking distribution* (CRD). Namely, a (parsimonious) model $P^*_{\mathcal{P}} = \sum_{\mathcal{C} \in \mathcal{P}} \alpha_{\mathcal{C}} \delta_{\sigma^*_{\mathcal{C}}}$ of a mixture of Dirac measures centred on local medians $\sigma^*_{\mathcal{C}}$ (*i.e.* Kemeny medians of Σ's conditional distribution given $\Sigma \in \mathcal{C}$ with $\mathcal{C} \in \mathcal{P}$) that approximates the ranking distribution P of interest better than a simple point mass δ_{σ^*} at a global median σ^* and with much less uncertainty than the raw empirical version $\widehat{P}_N = (1/N) \sum_{k \leq N} \delta_{\Sigma_k}$ of P based on $N << n!$ independent copies $\Sigma_1, \ldots, \Sigma_N$ of the r.v. Σ would. We highlight as main contributions:

- We demonstrate that the CRD $P^*_{\mathcal{P}}$ has attractive properties w.r.t. approximation of the distribution P, provided that $\mathcal{P}$ is chosen appropriately, when distortion is measured by the optimal transport metric whose cost function is the Kendall τ distance and prove a bound for the distortion that can be expressed using the local pairwise probabilities $\mathbb{P}\{\Sigma(i) < \Sigma(j) \mid \Sigma \in \mathcal{C}\}$ with $\mathcal{C} \in \mathcal{P}$ and $1 \leq i < j \leq n$ only (just like the local Kemeny medians in certain situations).
- We design a a statistical learning algorithm to recover, from ranking data, a CRD estimate balancing distortion bias against uncertainty/variability. These quantities can be accurately estimated when $\mathcal{P}$'s cells $\mathcal{C}$ are large enough, *i.e.* when the $\alpha_{\mathcal{C}}$'s are large enough. Precisely, a recursive partioning technique based on binary decision trees operating on $\mathfrak{S}_n$ (the root node) with splitting rules of the form '$\sigma(i) < \sigma(j)$' defined by pairs (i, j) of instances, we call COAST (COnsensus rAnking deciSion Trees).
- Beyond theoretical arguments, numerical experiments are presented, for the multi-criteria comparison of LLMs in particular, providing empirical evidence of the relevance of the COAST algorithm.

The article is structured as follows. In section 2, basic concepts and results pertaining to (statistical) consensus ranking theory are briefly recalled. The rigorous definition of consensus ranking distribution is given in section 3, together with the main (statistical) properties, while the COAST algorithm is introduced and (empirically) analyzed in section 4. Some concluding remarks are eventually collected in section 5. Technical details and additional results are deferred to the Appendix.

2 Background and preliminaries

We introduce the main notations and recall basic notions in consensus ranking theory for clarity's sake. The framework developed in the next section relies on these concepts and can be seen as a natural extension. 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 cardinality of any finite subset A by $\#A$, the support of a ranking distribution P by $\text{Supp}(P)$. Preferences expressed among a set of $n \geq 1$ items, indexed by $i \in [\![n]\!] := \{1, \ldots, n\}$ say, can be viewed (in the absence of ties for simplicity) as a permutation (full ranking) $\sigma \in \mathfrak{S}_n$ that maps any item i to its rank $\sigma(i)$. Certain marginals of the law P of a random permutation Σ on $\mathfrak{S}_n$ are of specific interest. In particular, for a pair of items $(i, j) \in [\![n]\!]^2$, the quantity $p_{i,j} = \mathbb{P}\{\Sigma(i) < \Sigma(j)\}$ for $\Sigma \sim P$ is referred to as a pairwise marginal of P and indicates the probability that item i is preferred to (ranked lower than) item j. Note that $p_{i,j} + p_{j,i} = 1$.

2.1 Consensus ranking and Kemeny medians

Consensus ranking, also referred to as *ranking aggregation*,is a problem than can be easily stated in an informal manner. Let $\sigma_1, \ldots, \sigma_N$ be a collection of $N \geq 1$ (full) rankings on $[\![n]\!]$. The goal is to find a ranking $\sigma^* \in \mathfrak{S}_n$ that summarizes it best. A rigorous and quantitative version of it can be formulated using a metric $d(.\,, .)$ on $\mathfrak{S}_n$ as the minimization problem:

$$\min_{\sigma \in \mathfrak{S}_n} \sum_{k=1}^{N} d(\sigma, \sigma_k), \tag{1}$$

Solutions to (1) always exist of course (since $\#\mathfrak{S}_n < +\infty$) but are not necessarily unique. Such barycentric permutations are called *consensus/median ranking*. Various distances have been considered in the consensus ranking literature, *e.g.* Spearman ρ, Spearman footrule, Hamming, see Chapter 11 in [9]. The most widely documented version of the problem is Kemeny ranking aggregation [22], that corresponds to the case where the metric chosen is the Kendall τ distance, *i.e.* the number of pairwise disagreements:

$$d_\tau(\sigma, \sigma') = \sum_{i<j} \mathbb{I}\{(\sigma(i) - \sigma(j))(\sigma'(i) - \sigma'(j)) < 0\}$$

for all $(\sigma, \sigma') \in \mathfrak{S}_n^2$. From a statistical learning perspective, (1) can be viewed as a M-estimation problem if one assumes that the rankings to be aggregated/summarized is composed of $N \geq 1$ independent copies $\Sigma_1, \ldots, \Sigma_N$ of a generic r.v. Σ, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and distributed according to an unknown probability distribution P on $\mathfrak{S}_n$ (*i.e.* $P(\sigma) = \mathbb{P}\{\Sigma = \sigma\}$ for all $\sigma \in \mathfrak{S}_n$). Similar to a median of a real valued r.v. Z, which is any scalar closest to Z in the L_1 sense, a (true) ranking median of distribution P w.r.t. a certain metric d on $\mathfrak{S}_n$ is any solution of

$$\min_{\sigma \in \mathfrak{S}_n} L_P(\sigma), \tag{2}$$

where the *ranking risk* $L_P(\sigma) = \mathbb{E}_{\Sigma \sim P}[d(\Sigma, \sigma)]$ is the expected distance between any permutation σ and the r.v. Σ. Statistical ranking aggregation consists in recovering a solution σ^* of this minimization problem, plus an estimate of this minimum $L_P^* = L_P(\sigma^*)$, as accurately as possible, from the training examples $\Sigma_1, \ldots, \Sigma_N$. A ranking median σ^* can be viewed as a location parameter, a central value, for the ranking distribution P, and the quantity L_P^* as a dispersion measure. Like (1), the minimization problem (2) has always a solution but can be multimodal. A major obstacle is of course that the ranking risk $L_P(.)$ is unknown, as the ranking distribution P. In absence of specific model assumptions about P (such as *e.g.* the Mallows model), when the ranking aggregation task is based on the Σ_k's only, the Empirical Risk Minimization (ERM) paradigm, see [36], encourages us to solve the statistical version of the problem obtained by replacing P in (2) with $\widehat{P}_N$: $\min_{\sigma \in \mathfrak{S}_n} \widehat{L}_N(\sigma)$ with

$$\widehat{L}_N(\sigma) = (1/N) \sum_{k=1}^{N} d(\Sigma_k, \sigma) = L_{\widehat{P}_N}(\sigma). \tag{3}$$

The performance of empirical minimizers $\widehat{\sigma}_N$ is studied in [24]: nonasymptotic bounds of order $O_{\mathbb{P}}(1/\sqrt{N})$ for the excess of risk $L_P(\widehat{\sigma}_N) - L_P^*$ in probability/expectation have been proved and shown to be sharp in the minimax sense, when $d = d_\tau$. If ranking aggregation (1) is NP-hard in general, see *e.g.* [17], exact solutions, called *Kemeny medians*, can be explicited in the Kendall τ case, when the pairwise probabilities $p_{i,j}$, $1 \leq i \neq j \leq n$, fulfill the *stochastic transitivity* property.

Definition 1. Let P be a probability law on $\mathfrak{S}_n$. It is said to be (weakly) stochastically transitive iff $\forall (i, j, k) \in [\![n]\!]^3$, $p_{i,j} \geq 1/2$ 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 $i < j$, P is said to be strictly stochastically transitive (SST).

Originally introduced in Psychology, see [7] and [11], these conditions were exploited more recently for ranking based on pairwise comparisons, refer to [34] or [33]). They are satisfied by the vast majority of popular parametric models such as Mallows [27] or Bradley-Terry-Luce-Plackett [30] models for instance. It is shown in [24] (see Theorem 5 therein) that, when stochastic transitivity is fulfilled, the ensemble of Kemeny medians is $\{\sigma \in \mathfrak{S}_n : \ (p_{i,j} - 1/2)(\sigma(j) - \sigma(i)) > 0 \text{ for all } i < j \text{ s.t. } p_{i,j} \neq 1/2\}$, and the related dispersion measure, denoted by

$$V_P := L_P^* \tag{4}$$

for clarity, is given by $V_P = \sum_{i<j} \min\{p_{i,j}, 1 - p_{i,j}\} = \sum_{i<j}\{1/2 - |p_{i,j} - 1/2|\}$.

Remark 1. (RANKING VARIABILITY) Except in the situation recalled above, calculating L_P^* is difficult. Since an expectation is generally easier to calculate/approximate than a minimum, an alternative dispersion measure is given by:

$$V_P' = \mathbb{E}[d(\Sigma, \Sigma')]/2, \tag{5}$$

where Σ' means an independent copy of Σ. Note that $V_P' \leq V_P \leq 2V_P'$ and, when $d = d_\tau$, $V_P' = \sum_{i<j} p_{i,j}(1 - p_{i,j})$.

2.2 Beyond ranking medians

Although very popular, once resolved, the consensus problem offers only a very crude summary of a ranking distribution. Other quantities have recently been defined in an attempt to describe certain properties of a ranking distribution more exhaustively, and issues relating to their statistical estimation have been explored. Because they easily relate to the concept we develop in the following section as will be seen later, we mention two such notions here.

Ranking depth. Ranking medians should be naturally considered as measures of central tendency in the ranking context. Following on from this idea, the concept of *ranking depth* has been introduced in [15]:

$$\forall \sigma \in \mathfrak{S}_n, \ \ D_P(\sigma) = \max_{(\sigma_1,\sigma_2)\in\mathfrak{S}_n^2} d(\sigma_1, \sigma_2) - L_P(\sigma).$$

Equipped with this notion for quantifying the centrality of any $\sigma \in \mathfrak{S}_n$, ranking medians are the deepest permutations and a center-outward ordering on the symmetric group is defined, which allows the concepts of quantiles, ranks and the powerful statistical methods in multivariate data analysis based on them to be directly extended to ranking data analysis. Refer also to [14] for applications to robust ranking.

Bucket ranking distribution. As noted in [1], $L_P(\sigma)$ can be seen as a distance between the ranking distributions P and δ_σ when considering the Wasserstein metric defined below.

Definition 2. Let $d : \mathfrak{S}_n^2 \to \mathbb{R}_+$ be a distance on $\mathfrak{S}_n$. The Wasserstein metric with d as cost function between two probability distributions P and P' on $\mathfrak{S}_n$ is given by:

$$W_d\left(P, P'\right) = \inf_{\Sigma\sim P,\ \Sigma'\sim P'} \mathbb{E}\left[d(\Sigma, \Sigma')\right], \tag{6}$$

where the infimum is taken over all possible couplings[1] (Σ, Σ') of (P, P').

Consensus ranking can be thus viewed as a radical *dimensionality reduction* procedure, summarizing P by its closest Dirac measure $\delta_{\sigma_P^*}$ w.r.t. to the probability metric above leading to the distortion $L_P^* = W_d(P, \delta_{\sigma_P^*})$. In [1], a more general framework for dimensionality reduction has been developed, which can be seen as an extension of consensus ranking. Under the initial hypothesis that the set $[\![n]\!]$ of instances to be ranked can be partitioned into *buckets* such that items in a certain bucket are either all ranked higher or else all ranked lower than items in another bucket with high probability, it permits to define a specific low dimensional approximation $\tilde{P}$ of a ranking distribution P, referred to as a *bucket ranking distribution*. Special attention is paid to the situation where the distortion is measured by the Kendall Wasserstein metric $W_{d_\tau}(\tilde{P}, P)$, since the minimum distortion can be then expressed as a function of the pairwise probabilities $p_{i,j}$ only (see Proposition 5 in [1]) and bucket ranking distributions with small distortion can be directly built by 'randomizing' empirical Kemeny medians, see section 3 in [1] for further details.

[1] By coupling of two probability distributions Q and Q' is meant a random pair (U, U') s.t. the marginal distributions of U and U' are Q and Q'.

The concepts mentioned above are relevant in the unimodal case especially. The definitions and algorithms presented and analyzed in the next sections are part of this line of research, with the notable exception that they relate to multimodal and/or spread-off ranking distributions P and crucially require the introduction of a notion of localization.

3 Consensus ranking distributions

In this section we develop a framework to approximate a possibly multimodal ranking distribution P, whose variability L_P^* is too large for an approximation by a simple Dirac mass $\delta_{\sigma_P^*}$ at a median to be satisfactory. After introducing the concept of consensus (ranking) distribution (CRD in abbreviated form) and describing the inherent trade-off between dimension reduction and distortion minimization, we then study the performance of statistical versions of CRD's from the perspective of empirical risk minimization.

3.1 Local ranking medians and partitions

The statistical recovery of a ranking distribution P based on an i.i.d. sample of $N \geq 1$ observations $\Sigma_1, \ldots, \Sigma_N$ is challenging because its dimensionality is $n! - 1$ in absence of restrictive parametric model assumptions: $P = \sum_{\sigma \in \mathfrak{S}_n} P(\sigma)\delta_\sigma$ with $\sum_{\sigma \in \mathfrak{S}_n} P(\sigma) = +1$. Since $n!$ is generally very large compared to the ranking dataset size N in practice, the raw empirical distribution $\widehat{P}_N = \sum_{\sigma \in \mathfrak{S}_n} \widehat{P}_N(\sigma)\delta_\sigma$ does not provide an accurate estimation of P. To avoid the rigidity of the approach consisting of specifying a statistical model with a low dimension relative to $n! - 1$ (typically of order n such as the Mallows or Plackett-Luce models, see [29]), we develop an alternative inspired by *local averaging* techniques used for nonparametric statistics of multivariate data. The approximation/estimation methodology we propose is based on the localization of the ranking median concept defined below.

Definition 3. (LOCAL CONSENSUS) Let Σ be a random permutation of $[\![n]\!]$ with law P and $\mathcal{C} \subset \mathfrak{S}_n$ s.t. $P(\mathcal{C}) > 0$. Denote Σ's conditional distribution given $\Sigma \in \mathcal{C}$ by $P_{\mathcal{C}}$. By local median $\sigma_{\mathcal{C}}^* \in \mathfrak{S}_n$ of the distribution P on region $\mathcal{C}$ is meant any ranking median of $P_{\mathcal{C}}$: $\mathbb{E}[d(\Sigma, \sigma_{\mathcal{C}}^*) \mid \Sigma \in \mathcal{C}] = V(\mathcal{C})$, denoting P's local variability on $\mathcal{C}$ by $V(\mathcal{C}) := \min_{\sigma \in \mathfrak{S}_n} L_{P_{\mathcal{C}}}(\sigma) = V_{P_{\mathcal{C}}}$.

We therefore propose defining a probabilistic notion of consensus, we call consensus ranking distribution, as a convex combination of Dirac masses at local medians on regions forming a partition of $\mathfrak{S}_n$.

Definition 4. (CONSENSUS DISTRIBUTION) Let $\mathcal{P}$ be a partition[2] of $\mathfrak{S}_n$ and P be a ranking distribution. A consensus distribution of P relative to $\mathcal{P}$ is any ranking distribution of the form

$$P_{\mathcal{P}} = \sum_{\mathcal{C} \in \mathcal{P}} P(\mathcal{C})\delta_{\sigma_{\mathcal{C}}^*}, \tag{7}$$

where $\sigma_{\mathcal{C}}^*$ is a local median of P on $\mathcal{C} \in \mathcal{P}$.

The approximant (7) is a generalization of the two 'extreme' cases constituted by the Dirac mass $P(\mathfrak{S}_n)\delta_{\sigma_P^*}$ at a median, which is of minimal dimension but has maximum distortion, and by the discrete law $P = \sum_{\sigma \in \mathfrak{S}_n} P(\sigma)\delta_\sigma$ itself which has zero distortion but is of dimension $n! - 1$. The choice of a consensus distribution is therefore a trade-off: the partition $\mathcal{P}$ must be fine enough so that the distortion of $P_{\mathcal{P}}$ is sufficiently small, while at the same time its cells $\mathcal{C}$ must be large enough for the statistics $\widehat{P}_N(\mathcal{C})$ based on $N \geq 1$ observations to be accurate estimates of $P(\mathcal{C})$, resulting in a reduced dimensionality of (7). The result below illustrates the balance, by providing a control of the distortion of (7) by the intra-cell ranking variability.

Proposition 1. (DISTORTION BOUND) *Let $\mathcal{P}$ be any partition of $\mathfrak{S}_n$ s.t. $P(\mathcal{C}) > 0$ for all $\mathcal{C} \in \mathcal{P}$. We have:*

$$W_d(P, P_{\mathcal{P}}) \leq \mathcal{E}_P(\mathcal{P}) := \sum_{\mathcal{C} \in \mathcal{P}} P(\mathcal{C})V(\mathcal{C}) \leq 2\mathcal{E}'_P(\mathcal{P}) := 2\sum_{\mathcal{C} \in \mathcal{P}} P(\mathcal{C})V'(\mathcal{C}), \tag{8}$$

where we set $V'(\mathcal{C}) = V'_{P_{\mathcal{C}}}$.

[2] By partition $\mathcal{P}$ of a set E is meant any collection of non empty pairwise disjoint subsets $\mathcal{C}$ of E s.t. $\cup_{\mathcal{C} \in \mathcal{P}} \mathcal{C} = E$.

Proof. Consider the coupling $(\Sigma, \Sigma_{\mathcal{P}})$ of $(P, P_{\mathcal{P}})$ defined by $\mathbb{P}\{\Sigma_{\mathcal{P}} = \sigma^*_{\mathcal{C}} \mid \Sigma \in \mathcal{C}\} = 1$ for all $\mathcal{C} \in \mathcal{P}$. The distortion can be then bounded as follows $W_\tau(P_{\mathcal{P}}, P) \leq \mathbb{E}[d(\Sigma, \Sigma_{\mathcal{P}})] = \sum_{\mathcal{C}\in\mathcal{P}} P(\mathcal{C})V(\mathcal{C})$, using the total probability rule and the second bound is simply due to the fact that $V(\mathcal{C}) \leq V'(\mathcal{C})$ for all $\mathcal{C} \in \mathcal{P}$. □

The coupling used in the argument above is not always optimal, meaning that (8) may be strict, as shown in the Supplementary Material (see the counterexample therein). However, there is equality in the case of the least fine partition $\mathcal{P}_1 := \{\mathfrak{S}_n\}$ and in the case of the finest partition $\mathcal{P}_{n!} := \cup_{\sigma\in \text{Supp}(P)}\{\sigma\}$. The same argument can be used to show that the difference between P's variability and that of its CRD $P_{\mathcal{P}}$ is controlled by the distortion, namely $|V_{\mathcal{P}} - V_{P_{\mathcal{P}}}| \leq W_d(P, P_{\mathcal{P}})$, and that the depth of any ranking median $\sigma^*_{P_{\mathcal{P}}}$ w.r.t. P is close to the maximum depth when the distortion is small: $D_P(\sigma^*_{P_{\mathcal{P}}}) \geq \max_{\sigma\in\mathfrak{S}_n} D_P(\sigma) - W_d(P, P_{\mathcal{P}})$. Observe also that

$$\mathcal{E}_P(\mathcal{P}) \leq \min_{\sigma\in\mathfrak{S}_n} \left\{ \sum_{\mathcal{C}\in\mathcal{P}} P(\mathcal{C})\mathbb{E}_{P_{\mathcal{C}}}[d(\sigma, \Sigma)] \right\} = V_P, \tag{9}$$

and, similarly, $\mathcal{E}'_P(\mathcal{P}) \leq V'_P$ for any partition $\mathcal{P}$. More generally, the finer the partition, the smaller the intra-cell variability, we have: $\mathcal{E}_P(\mathcal{P}') \leq \mathcal{E}_P(\mathcal{P})$ (respectively, $\mathcal{E}'_P(\mathcal{P}') \leq \mathcal{E}'_P(\mathcal{P})$) as soon as $\mathcal{P}'$ is a subpartition of $\mathcal{P}$, *i.e.* $\mathcal{P}'$ is a partition of $\mathfrak{S}_n$ such that, for all $\mathcal{C}' \in \mathcal{P}'$, there exists $\mathcal{C} \in \mathcal{P}$ s.t. $\mathcal{C}' \subset \mathcal{C}$. Of course, to adapt to the properties of the ranking distribution P, the partition $\mathcal{P}$ defining the CRD $P_{\mathcal{P}}$ must be learned from the data $\Sigma_1, \ldots, \Sigma_N$. Because it is easy to compute a statistical counterpart of $\mathcal{E}'(\mathcal{P})$ based on the Σ_k's, the bound in Proposition 1 permit to choose empirically a partition of controlled cardinality yielding a CRD with guarantees regarding distortion in the form of confidence bounds. It is the purpose of the subsection below to establish such generalization bounds.

Remark 2. (NONPARAMETRIC INSPIRATION) While the framework we develop here is based on partioning techniques just like histogram methods in multivariate data analysis, other approaches to statistical analysis of ranking data have been directly inspired by nonparametric techniques tailored to observations in $\mathbb{R}^q$ with $q \geq 1$. One may refer to *e.g.* [28] for the development of a method analogous to kernel smoothing for ranking distribution inference and to [17], [23] or [35] for "projection-like" approaches.

3.2 Learning CRD's - Statistical guarantees

Based on Proposition 1, we now study the principle of choosing the partition characterizing the CRD (7) by minimizing a statistical version of the quantity $\mathcal{E}'(\mathcal{P})$ over a finite collection Π_K of partitions $\mathcal{P}$ of $\mathfrak{S}_n$ with $K \leq n!$ cells $\mathcal{C}$ s.t. $P(\mathcal{C}) > 0$ from the perspective of generalization. Observe first that, based on the observations $\Sigma_1, \ldots, \Sigma_N$ supposedly available, a natural empirical counterpart of $\mathcal{E}'(\mathcal{P})$ is obtained in a plug-in fashion by replacing P with $\widehat{P}_N$ and is given by:

$$\widehat{\mathcal{E}}'_N(\mathcal{P}) = \sum_{\mathcal{C}\in\mathcal{P}} \widehat{P}_N(\mathcal{C})\widehat{V}'_N(\mathcal{C}), \tag{10}$$

where the local ranking variability estimates are

$$\widehat{V}'_N(\mathcal{C}) = \frac{2}{N_{\mathcal{C}}(N_{\mathcal{C}} - 1)} \sum_{i<j} \mathbb{I}\{(\Sigma_i, \Sigma_j) \in \mathcal{C}^2\} d(\Sigma_i, \Sigma_j),$$

with $N_{\mathcal{C}} = \#\{i \in \{1, \ldots, N\} : \Sigma_i \in \mathcal{C}\}$ and the convention that $\widehat{V}'_N(\mathcal{C}) = 0$ when $N_{\mathcal{C}} \leq 1$, where $\mathcal{C} \in \mathcal{P}$. The fluctuations of the stochastic process $\{\widehat{\mathcal{E}}'_N(\mathcal{P}) - \mathcal{E}(\mathcal{P}) : \mathcal{P} \in \Pi_K\}$ can be controlled by observing that: $\forall \mathcal{P} \in \Pi_K$,

$$\widehat{\mathcal{E}}'_N(\mathcal{P}) - \mathcal{E}(\mathcal{P}) = \left(\widehat{\mathcal{E}}'_N(\mathcal{P}) - \widetilde{\mathcal{E}}'_N(\mathcal{P})\right) + \left(\widetilde{\mathcal{E}}'_N(\mathcal{P}) - \mathcal{E}(\mathcal{P})\right), \tag{11}$$

where we set

$$\widetilde{\mathcal{E}}'_N(\mathcal{P}) = \frac{2}{N(N-1)} \sum_{i<j} \Phi_{\mathcal{P}}(\Sigma_i, \Sigma_j), \quad \text{with} \quad \Phi_{\mathcal{P}}(\sigma, \sigma') = \frac{1}{2} \sum_{\mathcal{C}\in\mathcal{P}} \frac{\mathbb{I}\{(\sigma, \sigma') \in \mathcal{C}^2\}}{P(\mathcal{C})} \cdot d(\sigma, \sigma'). \tag{12}$$

Hence, the quantity (12) is a (non degenerate) U-statistic of degree 2 with symmetric kernel $\Phi_{\mathcal{P}}$, see [25]. The maximal deviations of the second term on the right hand side of (11) can be thus controlled by means of classical extensions of tail inequalities for i.i.d. averages to U-statistics. Since those of the first term can be easily shown to be of order $O_{\mathbb{P}}(1/\sqrt{N})$, we have the following result (see technical details in the Appendix).

Proposition 2. *Let $K \in \{1, \ldots, n!\}$ and Π_K be a collection of partitions of $\mathfrak{S}_n$ counting each K cells $\mathcal{C}$ s.t. $P(\mathcal{C}) \geq \chi(\Pi_K) > 0$, where $\chi(\Pi_K)$ is the lower bound of the probability mass of every cell. Consider any minimizer $\widehat{\mathcal{P}}_N$ of the empirical criterion* (10) *over Π_K. For all $\delta \in (0,1)$, we have with probability at least $1-\delta$:*

$$\frac{W_d(P, P_{\widehat{\mathcal{P}}_N})}{2} \leq \mathcal{E}'(\widehat{\mathcal{P}}_N) \leq \min_{\mathcal{P} \in \Pi_K} \mathcal{E}'(\mathcal{P}) + 2\sqrt{\frac{\gamma(\Pi_K)}{2\chi(\Pi_K)} \frac{\log(4\#\Pi_K/\delta)}{N}} + 2C\frac{K||d||_\infty}{\chi^2(\Pi_K)}\sqrt{\frac{\log(8K/\delta)}{2N}},$$

if $N \geq \max\{\log(8K/\delta), (1+\sqrt{\log(4K/\delta)/2})/\chi(\Pi_K)\}$, where we set $\gamma(\Pi_K) = \max\left\{d(\sigma,\sigma') : (\sigma,\sigma') \in \mathcal{C}^2, \mathcal{C} \in \mathcal{P}, \mathcal{P} \in \Pi_K\right\}$.

The bound above provides guarantees for the principle consisting in choosing the partition defining the CRD by minimizing (10) over all $\mathcal{P}$ in the collection Π_K considered. As shown in the Appendix, a similar guarantee holds true, under appropriate conditions, for statistical versions of $P_{\widehat{\mathcal{P}}_N}$, where local medians $\sigma^*_{\mathcal{C}}$ and weights $P(\mathcal{C})$ are replaced with their empirical counterparts, see Proposition 3 therein.

Several aspects of the approach analyzed above are worth discussing, more details can be found in the Appendix B. First, the geometry of the cells plays a major role on the (approximate, incremental) resolution of $\min_{\mathcal{P} \in \Pi} \widehat{\mathcal{E}}'_N(\mathcal{P})$, an issue we also address in the next section. In general, local median $\sigma^*_{\mathcal{C}}$ does not necessarily belong to $\mathcal{C}$. Partitions whose cells enforce this containment are also those for which the NP-hardness of Kemeny aggregation can be bypassed, by deferring the computation of local medians to cells of low intra-cell distortion. Next, the number K of cells, a natural proxy for CRD complexity, is selected by complexity-regularized empirical risk minimization, yielding a near-oracle distortion bound following section 8 of [2]. Lastly, we discuss the relation with mixtures and clustering: a CRD generalizes flat clustering by controlling distortion at the distribution level with learnability guarantees; alternative partitioning strategies inspired by clustering, in particular a k-medians approach on $\mathfrak{S}_n$, are also possible but at the cost of increasing the complexity.

4 Partitioning trees and CRD's

Here, we describe at length a data-driven strategy to build a nested collection of partitions, defining a sequence of CRD's of increasing complexity (*i.e.* involving an increasing number of local medians) based on decision trees and incremental minimization of the $\widehat{\mathcal{E}}'_N$ criterion, referred to as the COAST (COnsensus rAnking diStribution Tree) algorithm. Next we present numerical experiments revealing its capacity to detect the local modes of ranking distributions.

4.1 The COAST algorithm

Here we focus on CRDs defined by binary trees and the Kendall τ distance, for reasons that we now explain. First, top-down tree-based techniques are popular for building partitions recursively. To grow the binary tree, one starts with the root node, corresponding to the trivial partition $\mathcal{P}_1 = \{\mathfrak{S}_n\}$ and one first splits $\mathfrak{S}_n$ into two disjoint regions. They are next each divided in two and the process continues until a stopping condition holds. At each iteration, the terminal leaves of the current binary tree are in 1-to-1 correspondence with the cells of a partition of $\mathfrak{S}_n$. The main difficulty lies in choosing an appropriate *binary splitting rule* (to divide a region in two sub-regions) in order to construct a partition that is well suited to the properties of the distribution P of the random ranking Σ being analyzed. We propose splitting rules of the form '$\sigma(i) < \sigma(j)$' *vs* '$\sigma(i) > \sigma(j)$' determined by a pair (i,j) s.t. $1 \leq i < j \leq n$, referred to as a *splitting pair*. The pair (i,j) is used to split a cell $\mathcal{C}$ of the current partition of $\mathfrak{S}_n$ into $\mathcal{C}^{(0)}_{i,j} := \{\sigma \in \mathcal{C} : \ \sigma(i) < \sigma(j)\}$ and $\mathcal{C}^{(1)}_{i,j} := \mathcal{C} \setminus \mathcal{C}^{(0)}_{i,j}$.

Each cell $\mathcal{C}$ of a partition $\mathcal{P}$ constructed this way defines a partial order on $\mathfrak{S}_n$ (and a total order when $\mathcal{C}$ reduces to a singleton of course) as long as (i,j) satisfy transitivity. Note that any node/cell

The COAST Algorithm

Input: Ranking data $\Sigma_1, \ldots, \Sigma_N$, threshold $\epsilon \geq 0$, ranking aggregation algorithm $\mathcal{A}$

Initialize: The root node is the trivial partition $\mathcal{P} = \{\mathfrak{S}_n\}$ and $\mathcal{J}_{\mathfrak{S}_n} = \{(i,j) : 1 \leq i < j \leq n\}$.

Repeat: while $\exists \mathcal{C} \in \mathcal{P}$ with $\widehat{V}'_N(\mathcal{C}) > \epsilon$

- For each such $\mathcal{C}$, choose $(i,j) = \arg\min_{(i,j)\in\mathcal{J}_\mathcal{C}} \widehat{P}_N(\mathcal{C}^{(0)}_{i,j})\widehat{V}'_N(\mathcal{C}^{(0)}_{i,j}) + \widehat{P}_N(\mathcal{C}^{(1)}_{i,j})\widehat{V}'_N(\mathcal{C}^{(1)}_{i,j})$
- Split $\mathcal{C}$ into $\mathcal{C}^{(0)}_{i,j}$ and $\mathcal{C}^{(1)}_{i,j}$, update $\mathcal{P}$ as $\{\mathcal{C} \notin \mathcal{P}(\epsilon)\} \cup \{\mathcal{C}^{(m)}_{i,j} : \ (\mathcal{C}, m) \in \mathcal{P}(\epsilon) \times \{0,1\}\}$

Local aggregation: for each $\mathcal{C} \in \mathcal{P}$, apply algorithm $\mathcal{A}$ to $\{\Sigma_1, \ldots, \Sigma_N\} \cap \mathcal{C}$ yielding $\widehat{\sigma}^*_\mathcal{C}$

Output: the empirical CRD $\widehat{P}_\mathcal{P} = \sum_{\mathcal{C}\in\mathcal{P}} \widehat{P}_N(\mathcal{C})\delta_{\widehat{\sigma}^*_\mathcal{C}}$

$\mathcal{C}$ of the tree/partition takes the form

$$\mathcal{C} = \{\sigma \in \mathfrak{S}_n : \ \sigma(i) < \sigma(j) \text{ for all } (i,j) \in \mathcal{I}\}, \tag{13}$$

where $\mathcal{I} \subset \{(i,j) : \ 1 \leq i \neq j \leq n\}$ is the set of 2-tuples used to build $\mathcal{C}$.

Another reason for choosing splitting rules based on pairwise comparisons is that, if the pairwise probabilities $p_{i,j}$ with $i < j$ do not characterize the ranking distribution P in general, the local versions $p_{i,j}(\mathcal{C}) = \mathbb{P}\{\Sigma(i) < \Sigma(j) \mid \Sigma \in \mathcal{C}\}$, with $\mathcal{C}$ of the form (13) and s.t. $P(\mathcal{C}) > 0$, do (see the Appendix for a rigorous formulation). In addition, when considering the notion of ranking variability based on the Kendall τ distance, the criterion $\mathcal{E}_P(\mathcal{P}) = \sum_{\mathcal{C}\in\mathcal{P}} P(\mathcal{C})V(\mathcal{C})$ (respectively, $\mathcal{E}'_P(\mathcal{P}) = \sum_{\mathcal{C}\in\mathcal{P}} P(\mathcal{C})V'(\mathcal{C})$) can be bounded as a function of such quantities. More generally, we have, for any $\mathcal{C} \subset \mathfrak{S}_n$ s.t. $P(\mathcal{C}) > 0$, $V(\mathcal{C})/2 \leq V'(\mathcal{C}) = \sum_{i<j} p_{i,j}(\mathcal{C})(1-p_{i,j}(\mathcal{C}))$. By convention, set $V(\mathcal{C}) = V'(\mathcal{C}) = V''(\mathcal{C}) = 0$ when $P(\mathcal{C}) = 0$. This allows us to define rules for choosing best the splitting pair to divide a cell $\mathcal{C}$ corresponding to an internal node of the tree: among the set $\mathcal{J}(\mathcal{C}) \subset \{(i,j) : 1 \leq i < j \leq n\}$ of *admissible* pairs, *i.e.* pairs that are compatible with the partial order defining $\mathcal{C}$, choose the splitting pair $(i_\mathcal{C}, j_\mathcal{C})$ to be a pair (i,j) minimizing the local dispersion criterion

$$\mathcal{E}'_{(i,j)}(\mathcal{C}) := P(\mathcal{C}^{(0)}_{i,j})V'(\mathcal{C}^{(0)}_{i,j}) + P(\mathcal{C}^{(1)}_{i,j})V'(\mathcal{C}^{(1)}_{i,j}). \tag{14}$$

Remark 3. (APPROXIMATE SPLITTING RULE) Another option for choosing the splitting pair $(i_\mathcal{C}, j_\mathcal{C})$ consists in selecting the pair (i,j) s.t. $|p_{i,j}(\mathcal{C}) - 1/2|$ is closest to 0, which also corresponds to the term which contributes most to the variability $V''(\mathcal{C})$ and leads to the most balanced child nodes possible. Compared to minimizing (14), this strategy has computational advantages, the complexity decreasing from $O((\#\mathcal{J}(\mathcal{C}))^4)$ to $O(\#\mathcal{J}(\mathcal{C}))$. Although, as our experiments have shown, it often produces a comparable reduction in distortion/variability, finding the conditions under which such splits are quasi-optimal remains an open question.

Pruning and CRD model selection. The partition $\mathcal{P}$ output by COAST is s.t. $\widehat{\mathcal{E}}'_N(\mathcal{P}) \leq \epsilon$. The tuning parameter $\epsilon \geq 0$ governs the tree size and therefore the CRD's complexity. If $\epsilon \geq \widehat{V}'_N$, the tree is reduced to its root node $\mathcal{T}_1 = \{\mathfrak{S}_n\}$ and the CRD estimate to a global (empirical) median, whereas, if $\epsilon = 0$, the CRD estimate coincides with the raw empirical distribution $\widehat{P}_N$. Since it can be difficult to choose ϵ, one strategy could be to select a 'small' value in order to first grow a 'large' tree, $\mathcal{T}_K$ with $K \leq N$ nodes say, then prune it, as follows. One starts from the 'large' tree $\mathcal{T}_K$ and form a sequence of subtrees $\mathcal{T}_K \supset \mathcal{T}_{K-1} \supset \ldots \supset \mathcal{T}_1$, defining a sequence of partitions $\mathcal{P}'_K, \ldots, \mathcal{P}'_1$ (note that $\#\mathcal{P}'_k = k$, $\mathcal{P}'_K = \mathcal{P}_K$ and $\mathcal{P}'_1 = \mathcal{P}_1$), by successively merging the internal nodes descending from the same parent, producing the smallest per-node increase in the variability $\widehat{\mathcal{E}}'_N$. One then selects the least complex subtree minimizing an estimate of the $\mathcal{E}'$ criterion. It can also be shown that this boils down to selecting the partition $\mathcal{P}'_k$ that minimizes the penalized criterion $\widehat{\mathcal{E}}'_N(\mathcal{P}'_k) + \lambda \cdot k$, for a specific value of $\lambda \geq 0$, a tuning parameter that governs the trade-off between empirical distortion and the number of cells (or local medians) retained for the approximation. See [3] for further details.

4.2 Numerical experiments and applications

We now illustrate the performance of COAST on both real and synthetic data, together with several applications of CRDs. Additional experiments are deferred to the Appendix, where we consider the

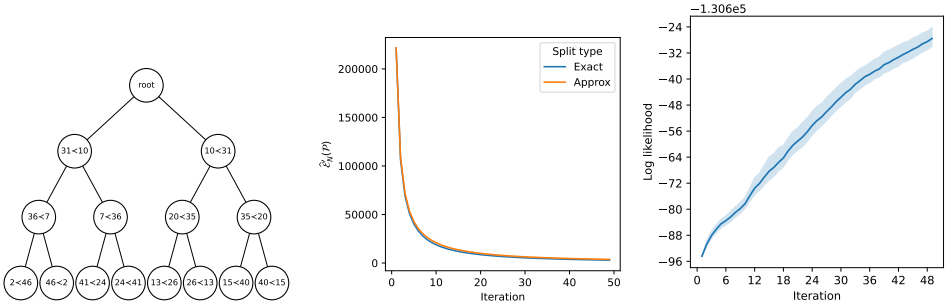


Figure 1: Experiments on estimation quality (a,b) and smoothing (c).

Sushi dataset [21] and synthetic Mallows and Plackett-Luce mixtures with up to $n = 150$ items, and further applications to anomaly detection. Full code and experiments at `https://github.com/ekhiru/COAST`.

The following experiments illustrate COAST on the TAC-08 benchmark [5], a real-world dataset for LLM benchmarking across a collection of tasks. Each of the $m = 930$ tasks induces a ranking of the $n = 48$ LLM systems by ordering them according to their performance for the given task. As different tasks rank systems differently, the resulting distribution over rankings is naturally multimodal and clusters around distinct performance profiles, making a single, global Kemeny median an inadequate summary.

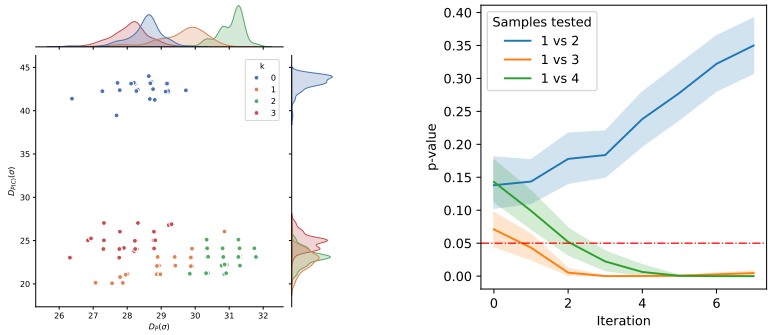
The ranking aggregation algorithm, denoted $\mathcal{A}$ in COAST, is Borda [8]. This choice is elaborated in Appendix B and summarized here: In COAST, the local median problem is solved only on low-variability cells since $\widehat{V}'_N(\mathcal{C}) < \epsilon$ triggers its termination. Borda, in this regime, provides a good approximation of the Kemeny median [4].

Benchmarking LLMs, visualization. Figure 1(a) shows the first three levels of the tree produced by the exact splitting rule of COAST ; the full tree is in Appendix B.9. The first split is on the pair $(i, j) = (31, 10)$: the single pairwise comparison that best separates the empirical distribution into two low-distortion sub-populations is "system 31 beats system 10" versus its converse. In other words, instances where system 31 outperforms system 10 exhibit pairwise behavior on the remaining systems that differs substantially from instances where the opposite holds. Subsequent splits admit an analogous interpretation, yielding an interpretable hierarchical decomposition a global median would fail to capture.

Splitting rule: Exact vs. Approximate. We compare on the same data the exact with the greedy rule of Remark 3. Figure 1(b) reports the empirical distortion $\widehat{\mathcal{E}}'_N$ along iterations for both: the curves are close, confirming that the greedy rule approximates the exact rule in practice despite not maximizing per-iteration decrease. The clear plateau provides a natural empirical stopping criterion. Appendices B.4 and B.4.1 further compare the criteria for rankings up to $n = 150$ and very noisy settings. Theoretical complexity differences are nevertheless visible at scale: for $n = 150$ and $K = 8$ the exact rule takes more than 15 minutes, against roughly 15 seconds for the greedy rule, at no measurable loss of distortion (see Table 3 in the Appendix).

Ranking distribution inference (smoothing). As recalled in the Introduction, most inference techniques fit a rigid statistical model [29], the nonparametric alternative of using $\widehat{P}_N$ being inevitably inaccurate when $N \ll n!$, see Sect. 3.1. Up to an affine transform, the local Kendall τ depth can be used to estimate a unimodal local distribution $P_{\mathcal{C}}$ via $\widetilde{P}_{\mathcal{C}}(\sigma) = Z_{\mathcal{C}}^{-1} \sum_{i<j} p_{\sigma^{-1}(i),\sigma^{-1}(j)}(\mathcal{C})$, $\sigma \in \mathcal{C}$, with $Z_{\mathcal{C}}$ computable in $O(n^2 \cdot \text{depth}(\mathcal{C}))$ (Appendix B.5). This results is a nonparametric, smoothed estimator of the whole ranking distribution. The approximation preserves essential properties such as median recovery. Figure 1(c) reports the log-likelihood of the resulting estimator on TAC-08 instance across iterations; it increases steadily and supports the same stopping choice as the distortion curve in panel (b).

DD-plots: local vs. global depth. Following [15], we extend DD-plots to the CRD setting. We sample $N = 400$ rankings from a mixture of $K = 4$ Mallows components with randomly chosen centers and dispersion $\phi = 0.1$. Figure 2(a) plots each ranking σ at coordinates $\big(D_{\widehat{P}_{\mathcal{C}}}(\sigma), D_{\widehat{P}_N}(\sigma)\big)$,



(a) (b)

Figure 2: (a) DD-plots illustrating local and global depths. (b) Homogeneity testing for the four distributions of Table 1.

i.e., its global depth against its local depth relative to the mixture component labeled 0. Global depth alone cannot tell the groups apart: along the vertical axis, points coming from different components overlap. Local depth, in contrast, separates them into clearly distinct clusters. The deepest ranking along the local axis, i.e., the local median $\widehat{\sigma}^*_{\mathcal{C}}$ of the mixture component labeled as 0, is not the global median, which is the deepest ranking along the global axis. A single global consensus therefore hides the heterogeneity of the distribution, and local depths reveal it. Analogous plots for other parametric models are given in Appendix B.6.

Homogeneity testing. Depths, both local and global, can also be employed to provide formal inference. We illustrate this through a nonparametric test of homogeneity between two Mallows model mixtures with $n = 6$. Four mixtures are considered, each with varying numbers of components K and scale parameters ϕ (see Table in Fig. 2). We perform the Wilcoxon rank-sum test comparing distribution 1 with each of the remaining distributions (i.e., 1 vs. 2, 1 vs. 3, and 1 vs. 4).

Distr.	ϕ	K
1	0.2	8
2	0.2	8
3	0.05	8
4	0.2	4

Table 1: Mixture parameters.

The Wilcoxon rank-sum test is performed using a reference sample of size 100 drawn from all distributions. The results, averaged over 200 repetitions, are shown in Figure 2. The p-values effectively detect differences between distributions when they exist, providing a formal inference that complements the visualization offered by ranking DD-plots. In particular, during the first iteration of the algorithm, the mixture P is effectively very uniform, and the resulting samples are consistently evaluated as different. However, when the distributions differ in K or the scale parameter ϕ, the test quickly identifies these differences, producing p-values close to zero by the second or third iteration. Remarkably, this procedure does not rely on any parametric assumptions about the underlying ranking models.

The key takeaway is that ranking depths in a CRD can be used to construct goodness-of-fit statistics, allowing one to assess how well a particular ranking model explains a given dataset.

5 Conclusion and perspectives

In this article, we introduced the new concept of consensus ranking distribution, thereby extending the notion of ranking median, particularly well suited to multimodal ranking distributions. We showed that the relevance of the statistical summary it provides is a delicate trade-off between dimension reduction and distortion minimization and proposed a tree-based learning algorithm for constructing CRDs from ranking data that achieve a satisfactory balance. Numerical results were presented to illustrate the effectiveness of the method and its usefulness for statistical ranking data analysis is demonstrated by several applications, which will hopefully pave the way for many more.

Acknowledgments and Disclosure of Funding

This research was supported by the Hi! PARIS Cluster 2030 and COST Action CA22137.

References

[1] M. Achab, A. Korba, and S. Clémençon. Dimensionality reduction and (bucket) ranking: a mass transportation approach. In A. Garivier and S. Kale, editors, *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98, pages 1–30. PMLR, 2019.

[2] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

[3] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, 1984.

[4] I. Caragiannis, A. D. Procaccia, and N. Shah. When do noisy votes reveal the truth? *ACM Transactions on Economics and Computation*, 4(3), 2016. Special issue on selected papers from EC-13.

[5] P. Colombo, N. Noiry, E. Irurozki, and S. Clémençon. What are the best systems? new perspectives on nlp benchmarking. In *Advances in Neural Information Processing Systems*, 2022.

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

[7] D. Davidson and J. Marschak. Experimental tests of a stochastic decision theory. *Measurement: Definitions and theories*, 17:274, 1959.

[8] J.-C. de Borda. Mémoire sur les élections au scrutin. *Histoire de l'Académie Royale des Sciences*, 1781.

[9] M. M. Deza and E. Deza. *Encyclopedia of Distances*. Springer, 2009.

[10] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, pages 613–622. ACM, 2001.

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

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

[13] Georg, J. C. B. Hannah, and Schollmeyer. Statistical models for partial orders based on data depth and formal concept analysis. In Inés, M. Jesús, Ślezak Dominik, P. Davide, B.-M. Bernadette, Y. R. R. C. Davide, and Couso, editors, *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 17–30. Springer International Publishing, 2022.

[14] M. Goibert, C. Calauzènes, E. Irurozki, and S. Clémençon. Robust consensus in ranking data analysis: Definitions, properties and computational issues. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 1–25. PMLR, 2023.

[15] M. Goibert, S. Clémençon, E. Irurozki, and P. Mozharovskyi. Statistical depth functions for ranking distributions: Definitions, statistical learning and applications. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

[16] A. Himmi, E. Irurozki, N. Noiry, S. Clémençon, and P. Colombo. Towards more robust NLP system evaluation: Handling missing scores in benchmarks. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11759–11785, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics.

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

[18] E. Irurozki and S. Clémençon. Universal aggregation of permutations. In *Proceedings of DA2PL 2022: From Multiple-Criteria Decision Aid to Preference Learning*, 2022.

[19] M. Jang, S. Kim, C. Suh, and S. Oh. Top-k ranking from pairwise comparisons: When spectral ranking is optimal. *arXiv preprint*, 2016.

[20] Y. Jiao, A. Korba, and E. Sibony. Controlling the distance to a kemeny consensus without computing it. In *Proceedings of ICML 2016*, 2016.

[21] T. Kamishima. Nantonac collaborative filtering: Recommendation based on order responses. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 583–588, 2003.

[22] J. G. Kemeny. Mathematics without numbers. *Daedalus*, 88:571–591, 1959.

[23] R. Kondor and M. S. Barbosa. Ranking with kernels in Fourier space. In *The 23rd Conference on Learning Theory*, pages 451–463, 2010.

[24] A. Korba, S. Clémençon, and E. Sibony. A learning theory of ranking aggregation. In *Proceeding of AISTATS 2017*, 2017.

[25] A. J. Lee. *U-statistics: Theory and practice*. Marcel Dekker, Inc., New York, 1990.

[26] R. Y. Liu, J. M. Parelius, and K. Singh. Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *The Annals of Statistics*, 27(3):783–858, 1999.

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

[28] Y. Mao and G. Lebanon. Non-parametric modeling of partially ranked data. *Journal of Machine Learning Research*, 9:2401–2429, 2008.

[29] J. I. Marden. *Analyzing and Modeling Rank Data*. CRC Press, London, 1996.

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

[31] A. Procaccia and N. Shah. Optimal aggregation of uncertain preferences. In *AAAI*, pages 608–614, 2016.

[32] R. Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 1980.

[33] N. Shah and M. Wainwright. Simple, robust and optimal ranking from pairwise comparisons. *arXiv preprint arXiv:1512.08949*, 2015.

[34] N. B. Shah, S. Balakrishnan, A. Guntuboyina, and M. J. Wainright. Stochastically transitive models for pairwise comparisons: Statistical and computational issues. *arXiv preprint arXiv:1510.05610*, 2015.

[35] E. Sibony, S. Clémençon, and J. Jakubowicz. MRA-based statistical learning from incomplete rankings. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 1432–1441. PMLR, 2015.

[36] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Lecture Notes in Statistics. Springer, 2000.

Table of Contents

A Technical proofs **13**

A.1 Proof of Proposition 2 13

A.2 Ranking Distribution Characterization by Local Pairwise Marginals 16

A.3 Alternative Dispersion Criterion 16

B Discussion, additional experiments and applications **16**

B.1 Choosing the number K of local medians 16

B.2 Relation with mixtures, clustering and the k-medians alternative 17

B.3 Local medians: Cell geometry and bypassing NP-hardness 17

B.3.1 Aggregation methods 18

B.4 Further experiments on the COAST 19

B.4.1 Recovering the partitions in very noisy settings 20

B.5 Smoothing: depths are a good approximation for ranking distributions and CRD . . 21

B.6 Visualization beyond ranking Depth: DD-plots on local vs. global depth 22

B.7 Anomaly detection 23

B.8 Real-world data: Sushi dataset 24

B.9 LLM-benchmarking: TAC-08 25

C Limitations **25**

A Technical proofs

A.1 Proof of Proposition 2

Observe first that, with probability one,

$$W_d(P, P_{\widehat{P}_N})/2 \leq \mathcal{E}'(\widehat{\mathcal{P}}_N) \leq \min_{\mathcal{P}\in\Pi_K} \mathcal{E}'(\mathcal{P}) + 2\max_{\mathcal{P}\in\Pi_K} \left|\widetilde{\mathcal{E}}'_N(\mathcal{P}) - \mathcal{E}'(\mathcal{P})\right| + 2\max_{\mathcal{P}\in\Pi_K} \left|\widetilde{\mathcal{E}}'_N(\mathcal{P}) - \widehat{\mathcal{E}}'_N(\mathcal{P})\right|. \tag{15}$$

Lemma 1. *For any partition $\mathcal{P}$ of $\mathfrak{S}_n$ s.t. $P(\mathcal{C}) > 0$ for all $\mathcal{C} \in \mathcal{P}$, we have: $\forall t > 0$:*

$$\mathbb{P}\left\{\left|\widehat{\mathcal{E}}'_N(\mathcal{P}) - \mathcal{E}'(\mathcal{P})\right| \geq t\right\} \leq 2e^{-Nt^2/\kappa_{\mathcal{P}}^2}, \tag{16}$$

where $\kappa_{\mathcal{P}} := \max_{\mathcal{C}\in\mathcal{P}} \left\{\max_{(\sigma,\sigma')\in\mathcal{C}^2} d(\sigma,\sigma')/P(\mathcal{C})\right\}$.

Proof. The bound (16) is obtained directly by applying the version of Hoeffding's exponential deviation inequality for U-statistics (see Theorem A of section 5.6 in [32]) to $\widehat{\mathcal{E}}'_N(\mathcal{P})$, a U-statistic of degree 2 whose symmetric kernel satisfies: $\forall(\sigma,\sigma') \in \mathfrak{S}_n^2$, $0 \leq \Phi_{\mathcal{P}}(\sigma,\sigma') \leq \kappa_{\mathcal{P}}$. □

As $\kappa_{\mathcal{P}} \leq \gamma(\Pi_K)/(2\chi(\Pi_K))$ for all $\mathcal{P} \in \Pi_K$, the union bound combined with (16) implies that, for all $\delta \in (0,1)$, we have:

$$\max_{\mathcal{P}\in\Pi_K} \left|\widetilde{\mathcal{E}}'_N(\mathcal{P}) - \mathcal{E}'(\mathcal{P})\right| \leq \sqrt{\frac{\gamma(\Pi_K)}{2\chi(\Pi_K)} \frac{\log(2\#\Pi_K/\delta)}{N}}, \tag{17}$$

with probability larger than $1-\delta$.

Next, let us turn to the third term on the right hand side of (15). Combining Hoefdding's inequality with the union bound, we immediately obtain that the event $\cap_{\mathcal{C}\in\mathcal{P}}\{N_{\mathcal{C}} \geq 2\}$ occurs with overwhelming probability:

$$\mathbb{P}\left\{\cap_{\mathcal{C}\in\mathcal{P}}\{N_{\mathcal{C}} \geq 2\}\right\} \geq 1 - Ke^{-2(N\chi(\Pi_K)-1)^2}, \tag{18}$$

as soon as $N \geq 1/\chi(\Pi_K)$. On this event, notice that, for all $\mathcal{P} \in \Pi_K$, we have:

$$\left|\widetilde{\mathcal{E}}'_N(\mathcal{P}) - \widehat{\mathcal{E}}'_N(\mathcal{P})\right| = \left|\frac{2}{N(N-1)}\sum_{1\leq i<j\leq N}\sum_{\mathcal{C}\in\mathcal{P}}\left(\frac{N-1}{N_{\mathcal{C}}-1} - \frac{1}{P(\mathcal{C})}\right)\mathbb{I}\{(\Sigma_i,\Sigma_j)\in\mathcal{C}^2\}d(\Sigma_i,\Sigma_j)\right|$$
$$\leq K||d||_\infty \cdot \max_{\mathcal{C}\in\mathcal{P}}\left|\frac{N-1}{N_{\mathcal{C}}-1} - \frac{1}{P(\mathcal{C})}\right|. \tag{19}$$

Combining Hoefdding's inequality with the union bound, we immediately obtain that, for all $\delta \in (0,1)$, we have with probability at least $1-\delta$:

$$\max_{\mathcal{C}\in\mathcal{P}}\left|\widehat{P}_N(\mathcal{C}) - P(\mathcal{C})\right| \leq \sqrt{\frac{\log(2K/\delta)}{2N}}, \tag{20}$$

where $\widehat{P}_N(\mathcal{C}) = N_{\mathcal{C}}/N$ for all $\mathcal{C}\in\mathcal{P}$. Observing that $|(N_{\mathcal{C}}-1)/(N-1) - \widehat{P}_N(\mathcal{C})| \leq 2/(N-1)$ with probability one and combining the union bound with (18), (19), (20) and the Taylor expansion

$$\frac{1}{x} = \frac{1}{a} - \frac{(x-a)}{a^2} + \frac{(x-a)^2}{xa^2},$$

one obtains that, for all $\delta\in(0,1)$, we have with probability larger than $1-\delta$:

$$\left|\widetilde{\mathcal{E}}'_N(\mathcal{P}) - \widehat{\mathcal{E}}'_N(\mathcal{P})\right| \leq C\frac{K||d||_\infty}{\chi^2(\Pi_K)}\sqrt{\frac{\log(4K/\delta)}{2N}}, \tag{21}$$

where C is a universal constant, as soon as $N \geq \max\{\log(4K/\delta),\ (1+\sqrt{\log(2K/\delta)/2})/\chi(\Pi_K)\}$. Finally, combining (15), (27) and (21) with the union bound yiels the desired result.

Statistical Versions of Optimal CRDs

Let $1 \leq K \leq n!$ and Π_K be a collection of partitions of $\mathfrak{S}_n$ counting K cells. Consider now the distribution

$$\widehat{P}_{\widehat{\mathcal{P}}_N} := \sum_{\mathcal{C}\in\widehat{\mathcal{P}}_N}\widehat{P}_N(\mathcal{C})\delta_{\widehat{\sigma}^*_{\mathcal{C}}}, \tag{22}$$

where $\widehat{\mathcal{P}}_N$ is a minimizer of (10) over Π_K and, for all $\mathcal{C}\in\widehat{\mathcal{P}}_N$,

$$\widehat{\sigma}^*_{\mathcal{C}} \in \operatorname*{argmin}_{\sigma\in\mathfrak{S}_n}\sum_{i:\ \Sigma_i\in\mathcal{C}} d(\Sigma_i,\sigma). \tag{23}$$

Conditioned upon the Σ_i's, we introduce as well the ranking distribution

$$\widetilde{P}_{\widehat{\mathcal{P}}_N} := \sum_{\mathcal{C}\in\widehat{\mathcal{P}}_N}P(\mathcal{C})\delta_{\widehat{\sigma}^*_{\mathcal{C}}}, \tag{24}$$

By triangle inequality, we have:

$$W_d(P,\widehat{P}_{\widehat{\mathcal{P}}_N}) \leq W_d(\widetilde{P}_{\widehat{\mathcal{P}}_N},\widehat{P}_{\widehat{\mathcal{P}}_N}) + W_d(P,\widetilde{P}_{\widehat{\mathcal{P}}_N}). \tag{25}$$

Conditioned upon the Σ_i's, consider the coupling $(\widetilde{\Sigma},\Sigma)$ of the pair of ranking distributions $(\widetilde{P}_{\widehat{\mathcal{P}}_N},P)$ defined by:

$$\forall\mathcal{C}\in\widehat{\mathcal{P}}_N,\ \ \mathbb{P}\{\widetilde{\Sigma}\in\mathcal{C}\mid\Sigma\in\mathcal{C},\ \Sigma_1,\ \ldots,\ \Sigma_N\} = +1 \text{ almost-surely.} \tag{26}$$

In order to lighten the notation, we omit to write explicitely the conditioning upon the Σ_i's from now on. We have:

$$W_d(P, \widetilde{P}_{\widehat{\mathcal{P}}_N}) \leq \mathbb{E}[d(\widetilde{\Sigma}, \Sigma)] = \sum_{\mathcal{C} \in \widehat{\mathcal{P}}_N} P(\mathcal{C}) \mathbb{E}[d(\Sigma, \widehat{\sigma}^*_{\mathcal{C}}) \mid \Sigma \in \mathcal{C}] \leq$$
$$\mathcal{E}(\widehat{\mathcal{P}}_N) + \sum_{\mathcal{C} \in \widehat{\mathcal{P}}_N} P(\mathcal{C}) \left(\mathbb{E}[d(\Sigma, \widehat{\sigma}^*_{\mathcal{C}}) \mid \Sigma \in \mathcal{C}] - \mathbb{E}[d(\Sigma, \sigma^*_{\mathcal{C}}) \mid \Sigma \in \mathcal{C}] \right) \leq$$
$$2\mathcal{E}'(\widehat{\mathcal{P}}_N) + \max_{\mathcal{P} \in \Pi_K} \max_{\mathcal{C} \in \mathcal{P}} \left(\mathbb{E}[d(\Sigma, \widehat{\sigma}^*_{\mathcal{C}}) \mid \Sigma \in \mathcal{C}] - \mathbb{E}[d(\Sigma, \sigma^*_{\mathcal{C}}) \mid \Sigma \in \mathcal{C}] \right). \quad (27)$$

Each expectation $\mathbb{E}[d(\Sigma, \widehat{\sigma}^*_{\mathcal{C}}) \mid \Sigma \in \mathcal{C}]$ involved in the summation on the right hand side of (27) can be bounded by means of the lemma below, which corresponds to a version of Proposition 9 in [24] where the number of observations available to compute an empirical median of $P_{\mathcal{C}}$ is random, equal to $N_{\mathcal{C}}$, which is equivalent to $NP(C)$ with probability one.

Lemma 2. *Let $\mathcal{C} \subset \mathfrak{S}_n$ s.t. $P(\mathcal{C}) \geq \chi > 0$. Let $\widehat{\sigma}^*_{\mathcal{C}}$ be defined by* (23)*. Then, for all $\delta \in (0, 1)$, we have with probability larger than $1 - \delta$:*

$$\mathbb{E}[d(\Sigma, \widehat{\sigma}^*_{\mathcal{C}}) \mid \Sigma \in \mathcal{C}] - \mathbb{E}[d(\Sigma, \sigma^*_{\mathcal{C}}) \mid \Sigma \in \mathcal{C}] \leq n(n-1) \sqrt{\frac{\log(2n(n-1)/\delta)}{\chi N}}, \quad (28)$$

as soon as $N \geq \log(2/\delta)/2$.

Proof. It follows from Hoeffding's inequality that, with probability at least $1 - \delta$:

$$\widehat{P}_N(\mathcal{C}) \geq P(\mathcal{C}) - \sqrt{\frac{\log(1/\delta)}{2N}}.$$

Hence, with probability larger than $1 - \delta$, we have:

$$N_{\mathcal{C}} \geq N\chi/2,$$

as soon as $N \geq 2\log(1/\delta)/\chi$. Combining this with the bound stated in assertion (ii) of Proposition 9 in [24] (applied conditionally to $N_{\mathcal{C}}$) and the union bound yields the desired result. □

Applying Lemma 2 combined with the union bound, we obtain that, with probability at least $1 - \delta$:

$$\max_{\mathcal{P} \in \Pi_K} \max_{\mathcal{C} \in \mathcal{P}} \left(\mathbb{E}[d(\Sigma, \widehat{\sigma}^*_{\mathcal{C}}) \mid \Sigma \in \mathcal{C}] - \mathbb{E}[d(\Sigma, \sigma^*_{\mathcal{C}}) \mid \Sigma \in \mathcal{C}] \right) \leq n(n-1) \sqrt{\frac{\log(2n(n-1)K\#\Pi_K/\delta)}{\chi(\Pi_K)N}}, \quad (29)$$

as soon as $N \geq 2\log(2K\#\Pi_K/\delta)/\chi(\Pi_K)$.

Now turn to the first term on the right hand side of (25). Conditioned upon the Σ_i's, consider any coupling $(\widetilde{\Sigma}, \widehat{\Sigma})$ of the pair of ranking distributions $(\widetilde{P}_{\widehat{\mathcal{P}}_N}, \widetilde{P}_{\widehat{\mathcal{P}}_N})$ such that: $\forall \mathcal{C} \in \widehat{\mathcal{P}}_N$,

$$\begin{array}{lll} \mathbb{P}\{\widehat{\Sigma} \in \mathcal{C} \mid \widetilde{\Sigma} \in \mathcal{C},\ \Sigma_1,\ \ldots,\ \Sigma_N\} & = +1 \quad \text{almost-surely} & \text{if } P(\mathcal{C}) \leq \widehat{P}_N(\mathcal{C}), \\ \mathbb{P}\{\widetilde{\Sigma} \in \mathcal{C} \mid \widehat{\Sigma} \in \mathcal{C},\ \Sigma_1,\ \ldots,\ \Sigma_N\} & = +1 \quad \text{almost-surely} & \text{if } P(\mathcal{C}) \geq \widehat{P}_N(\mathcal{C}). \end{array}$$

We clearly have:

$$W_d(\widetilde{P}_{\widehat{\mathcal{P}}_N}, \widehat{P}_{\widehat{\mathcal{P}}_N}) \leq \mathbb{E}[d(\widetilde{\Sigma}, \widehat{\Sigma})] \leq ||d||_\infty \sum_{\mathcal{C} \in \widehat{\mathcal{P}}_N} \left| \widehat{P}_N(\mathcal{C}) - P(\mathcal{C}) \right| \leq ||d||_\infty K \max_{\mathcal{P} \in \Pi_K} \max_{\mathcal{C} \in \mathcal{P}} \left| \widehat{P}_N(\mathcal{C}) - P(\mathcal{C}) \right|. \quad (30)$$

Combined with (20) and the union bound, the bound above permits to show that, with probability larger than $1 - \delta$:

$$W_d(\widetilde{P}_{\widehat{\mathcal{P}}_N}, \widehat{P}_{\widehat{\mathcal{P}}_N}) \leq ||d||_\infty K \sqrt{\frac{\log(2K\#\Pi_K/\delta)}{2N}}. \quad (31)$$

From Proposition 2 combined with (25), (27), (29) and (31) we deduce the following result.

Proposition 3. *Suppose the assumptions of Proposition 2 are satisfied. Consider any minimizer $\widehat{\mathcal{P}}_N$ of the empirical criterion* (10) *over Π_K. For all $\delta \in (0,1)$, we have with probability at least $1-\delta$:*

$$W_d(P, \widehat{P}_{\widehat{\mathcal{P}}_N})/2 \leq \mathcal{E}'(\widehat{\mathcal{P}}_N) \leq \min_{\mathcal{P}\in\Pi_K} \mathcal{E}'(\mathcal{P}) + 2\sqrt{\frac{\gamma(\Pi_K)}{2\chi(\Pi_K)}\frac{\log(12\#\Pi_K/\delta)}{N}} + 2C\frac{K||d||_\infty}{\chi^2(\Pi_K)}\sqrt{\frac{\log(24K/\delta)}{2N}}$$
$$\binom{n}{2}\sqrt{\frac{3\log(2n(n-1)K\#\Pi_K/\delta)}{\chi(\Pi_K)N}} + \frac{||d||_\infty K}{2}\sqrt{\frac{\log(6K\#\Pi_K/\delta)}{2N}},$$

when $N \geq \max\{\log(24K/\delta),\ (1+\sqrt{\log(12K/\delta)/2})/\chi(\Pi_K),\ 2\log(6K\#\Pi_K/\delta)/\chi(\Pi_K)\}$.

A.2 Ranking Distribution Characterization by Local Pairwise Marginals

While it is well-known that the pairwise probabilities $p_{i,j}$ do not characterize a ranking distribution P in general, the collection of local pairwise probabilities $p_{i,j}(\mathcal{C})$, where $\mathcal{C}$ describes the ensemble of ranking subsets of the form (13). Consider the lexicographic order on $\{(i,j) \in [\![n]\!]^2 :\ i<j\}$: $(i,j) \prec (k,l)$ iff either '$i<k$' or else '$i=k$ and $j<l$'.

Equipped with this notation, for any $\sigma \in \mathfrak{S}_n$, one can write

$$P(\sigma) = \mathbb{P}\left\{\bigcap_{i<j}\{(\Sigma(i)-\Sigma(j))(\sigma(i)-\sigma(j))>0\}\right\}$$
$$= (p_{1,2}\mathbb{I}\{\sigma(1)<\sigma(2)\} + (1-p_{1,2})\mathbb{I}\{\sigma(2)<\sigma(1)\}) \times$$
$$\mathbb{P}\left\{\bigcap_{k<l:\ (1,2)\prec(k,l)}\{(\Sigma(k)-\Sigma(l))(\sigma(l)-\sigma(l))>0\} \mid (\Sigma(1)-\Sigma(2))(\sigma(1)-\sigma(2))>0\right\} \quad (32)$$

and the result follows by induction.

The assumption underlying the COAST algorithm is that it is possible to approximate the distribution P under study based on knowledge of the local pairwise probabilities $p_{i,j}(\mathcal{C})$, for a few well-chosen subsets $\mathcal{C}$ of the form (13) forming a partition of $\mathfrak{S}_n$.

A.3 Alternative Dispersion Criterion

The dispersion of a ranking probability distribution P can be measured by

$$V''_P = \sum_{k<l}\min\{p_{k,l},\ 1-p_{k,l}\},$$

which quantity can be directly computed from the pairwise probabilities $p_{i,j}$, as V'_P. Notice that $V_P \leq V'_P \leq V''_P$ and the equalities hold true when P is SST. Observe also that, for any partition $\mathcal{P}$ of $\mathfrak{S}_n$, we have:

$$\sum_{\mathcal{C}\in\mathcal{P}} P(\mathcal{C})V''(\mathcal{C}) \leq V''_P. \quad (33)$$

This simply results from the fact that: $\forall \mathcal{C} \in \mathcal{P}$,

$$V''(\mathcal{C}) = \sum_{k<l}\min\{p_{k,l}(\mathcal{C}),\ 1-p_{k,l}(\mathcal{C})\}$$

and, for all $k<l$,

$$p_{k,l} = \sum_{\mathcal{C}\in\mathcal{P}} P(\mathcal{C})p_{k,l}(\mathcal{C}). \quad (34)$$

B Discussion, additional experiments and applications

B.1 Choosing the number K of local medians

A crucial model selection issue remains to be solved: determining an appropriate value for the number $K \in \{1,\ \ldots,\ n!\}$ of local medians or cells involved in (7), a natural proxy for the 'complexity' of

the resulting CRD, so as to achieve a satisfactory 'distortion bias *vs* variance' trade-off. The popular idea of complexity regularization method consists in adding a penalty term $\text{PEN}(N, m)$ to $\mathcal{E}(\widehat{\mathcal{P}})$. Consider a sequence $\Pi_1, \ldots, \Pi_M$ of $M \geq 1$ collections of partitions of $\mathfrak{S}_n$, Π_m containing only partitions of cardinality m and let $\widehat{\mathcal{P}}_N^{(m)}$ be a minimizer of the empirical criterion (10) over Π_m. One then selects $\widehat{\mathcal{P}}_N^{(\widehat{m})}$ where the index $\widehat{m}$ minimizes $\widehat{\mathcal{E}}'(\widehat{\mathcal{P}}_N^{(m)}) + \text{PEN}(N, m)$ over $\{1, \ldots, M\}$ to avoid overfitting. A classic approach consists in choosing for PEN(N,M) an upper confidence bound for $\sup_{\mathcal{P} \in \Pi_M} |\widehat{\mathcal{E}}(\mathcal{P}) - \mathcal{E}(\mathcal{P})|$. A (deterministic) bound can be derived from the argument of Proposition 2 for instance. Following line by line the reasoning in section 8 of [2], it can be shown that the partition thus selected defines a CRD that nearly achieves the distortion that would be obtained with the help of an oracle, revealing the value m^* minimizing $\mathbb{E}[\mathcal{E}'(\widehat{\mathcal{P}}_N^{(m)})]$. Due to space limitations, details are left to the reader.

B.2 Relation with mixtures, clustering and the k-medians alternative

A CRD can also be interpreted as a nonparametric mixture model, where each component is a Dirac mass at a local median $\sigma_{\mathcal{C}}^*$ with weight $P(\mathcal{C})$. This contrasts with parametric mixture models such as Mallows or Plackett-Luce mixtures, which impose a specific within-component distribution. A CRD makes no such assumption: it only requires that rankings within each cell share a common median, making it robust to distributional misspecification.

A CRD generalizes clustering: Whereas standard unsupervised clustering finds a partition of the observed rankings with no statistical guarantees on the underlying distribution, CRD controls distortion at the distribution level and provides the learnability guarantees established in Section 3. The tree-based structure of COAST additionally supports pruning, see Section 4.

Nevertheless, the process of finding the partition can be replaced by a off-the-shelf clustering algorithm, a k-medians formulation on $\mathfrak{S}_n$ is a natural alternative: since the Kendall distance and local medians are well-defined on $\mathfrak{S}_n$, one can directly minimize $\mathcal{E}$ over partitions of cardinality k via alternating Voronoi reassignment and median recomputation, which also yields a monotone decrease of the empirical distortion. Here we note the key practical distinction: in COAST, the splitting criterion relies entirely on pairwise probabilities $\hat{p}_{i,j}(\mathcal{C})$, which are $O(n^2)$ to compute, and local medians are computed only once at the terminal leaves of the final tree, where low intra-cell distortion makes approximate aggregation methods provably accurate. Moreover, the geometry of the cells does not guarantee inclusiveness of the local median. A k-medians algorithm, by contrast, must recompute a local Kemeny median at every iteration and for every cell, including in high-dispersion regimes where exact Kemeny aggregation is NP-hard and approximations are less reliable. COAST thus sidesteps the computational bottleneck by design, deferring median computation to precisely the regime where it is tractable.

B.3 Local medians: Cell geometry and bypassing NP-hardness

The geometry of the cells $\mathcal{C}$ plays a central role in the practical tractability of CRD estimation. A local median $\sigma_{\mathcal{C}}^* \in \arg\min_{\sigma \in \mathfrak{S}_n} \sum_{\pi \in \mathcal{C}} d_\tau(\sigma, \pi)$ does not in general belong to the cell $\mathcal{C}$ itself, since the minimizer of the within-cell Kendall distance is defined over the full permutation space $\mathfrak{S}_n$. Partitions whose cells enforce the containment $\sigma_{\mathcal{C}}^* \in \mathcal{C}$ are of particular interest. Recall that $\mathcal{C}$ of the tree/partition takes the form

$$\mathcal{C} = \{\sigma \in \mathfrak{S}_n : \ \sigma(i) < \sigma(j) \text{ for all } (i,j) \in \mathcal{I}\},$$

where $\mathcal{I} \subset \{(i,j) : \ 1 \leq i \neq j \leq n\}$ is the set of 2-tuples used to build $\mathcal{C}$. For this choice of cell the local medians Borda and Kemeny clearly belong to $\mathcal{C}$.

In general, in a low-dispersion regime, rank aggregation procedures such as Borda count achieve approximation guarantees that make exact Kemeny computation unnecessary [4].

These observations form the foundation of the COAST strategy: defer the median computation to the terminal leaves of the partition tree, where within-cell dispersion is already small. At that stage, the NP-hardness of Kemeny aggregation can be effectively bypassed by using Borda, yielding provably similar results.

As a take-away: Rather than performing a single median computation at the root node (which is both computationally complex and likely to produce a non-representative median) COAST constructs K

cells and estimates one Borda median within each of the K cells. Because each cell has very low distortion, the resulting medians are both representative and easy to compute.

In the next lines, we analyze the behavior of local medians when conditioning on the partition structure.

Remark 4. The conditional distribution $P_{\mathcal{C}}^{(l)}$ is not SST in general. However, even when the pairs (i, j) are chosen randomly, in practice, provided that the pairs are consistent with the median, i.e., $P_{i,j} > .5$ the recovered median is a good approximation of the median, specially when the distribution is concentrated.

To illustrate these points, we include a randomized rank aggregation algorithm as an example. This algorithm, presented in [18], is essentially a Markov chain that starts from a random ranking and moves at each step to the neighboring ranking (at distance 1) with the largest depth.

Let us consider a sample $\{\sigma_1, \ldots, \sigma_m\}$ drawn from a SST distribution with center σ_0 and empirical measure $\widehat{P}$. In the plots in Figure 3, each point represents a ranking σ in the sample, placed at coordinates $(d(\sigma, \sigma_0), D_{\hat{P}}(\sigma))$. The goal of the algorithm is therefore to reach the top-left-most point, corresponding to σ_0.

The lines indicate the paths of the MCMC chain over 5 runs, which, by construction, always move "up" in depth. While the x-axis is not known to the algorithm in practice, as shown in [18], increasing depth $D_{\hat{P}}(\sigma)$ corresponds to decreasing in $d(\sigma, \sigma_0)$ (moving left) for all SST distributions.

Figure 3(a) illustrates the behavior of this Markovian algorithm on a distribution with a single SST component. Let $P^{(l)}$ be the distribution after conditioning l times. In panel (b) we consider the partition $\mathcal{C}$ that consists on recursively conditioning 3 times, always on the pairs $(i, j) = \arg\max P_{i,j}^{(l)}$ for $l \leq 3$.

Panels (c) and (d) repeat the process 6 and 9 times respectively. Note that, or $n = 6$, in at most 14 conditioning steps, the median becomes the only element in the partition $\mathcal{C}$. By construction, $D_{P_{\mathcal{C}}}$ is SST for every $\mathcal{C}$. Therefore, and for the same proof as in [18], the Markovian algorithm is guaranteed to find the median σ_0. The plots confirm this fact.

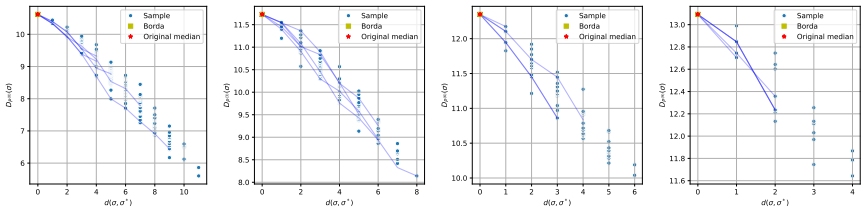


Figure 3: Markovian rank-aggregation algorithm for the case of a simple SST component. The sequence of pairs $(i, j) \in \mathcal{I}(\mathcal{C})$ is chosen such that $(i, j) = \arg\max P_{i,j}$, so that $P(\mathcal{C})$ is SST at every conditioning step of the algorithm. We include the results of conditioning 0, 3, 6 and 9 times for $n = 6$.

As pointed out above, when conditioning l times on arbitrary pairs, the resulting sequence $P^{(l)}$ is not SST in general. Nevertheless, the algorithm still performs well in practice. As shown in [18], the procedure is robust: even though not all depth-increasing paths converge to the local median, the majority of them do, allowing the median to be reliably recovered. See the illustration of the CRD of $K = 4$ on Figure 4.

B.3.1 Aggregation methods

We use Borda count as the local aggregation algorithm $\mathcal{A}$ within COAST's terminal cells, since in the low-dispersion regime where local aggregation is invoked, Borda closely approximates the Kemeny median [4] at $O(n \log n)$ cost per cell. We compare it against two other rules. Copeland ranks items by pairwise majority wins, $\sum_{j \neq i} \mathbb{I}\{p_{i,j}(\mathcal{C}) > 1/2\}$. It coincides with the Condorcet winner when one exists, and costs $O(n^2)$ per cell. Schulze ranks items via the strongest directed path in the majority graph weighted by $p_{i,j}(\mathcal{C})$, at $O(n^3)$ cost per cell matching COAST's own exact splitting rule.

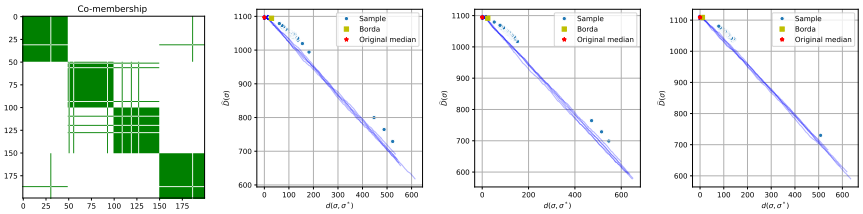


Figure 4: Markovian rank-aggregation algorithm for the case of a simple SST component. The sequence of pairs $(i, j) \in \mathcal{I}(\mathcal{C})$ is chosen such that $(i, j) = \arg\max P_{i,j}$, so that $P(\mathcal{C})$ is SST at every conditioning step of the algorithm. We include the results of conditioning 0, 3, 6 and 9 times for $n = 6$.

Comparison setup. We compare the three rules as candidates for $\mathcal{A}$ using the approximate splitting rule of Remark 3, across mixtures of $K \in \{4, 8\}$ components, drawn both from Mallows models with dispersion $\phi \in \{0.1, 0.3\}$ or from Plackett-Luce models with exponentially decreasing weights, for $n \in \{10, 50, 100, 150\}$. Each configuration was run 5 times. We report the error (Kendall τ distance) of each rule's local median estimate relative to the Borda local median.

n	K	Error Copeland	Error Schulze
10	4	0.475	0.450
	8	1.117	1.022
50	4	1.133	1.142
	8	0.906	0.900
100	4	3.292	3.275
	8	1.950	1.911
150	4	4.108	4.100
	8	3.072	3.028

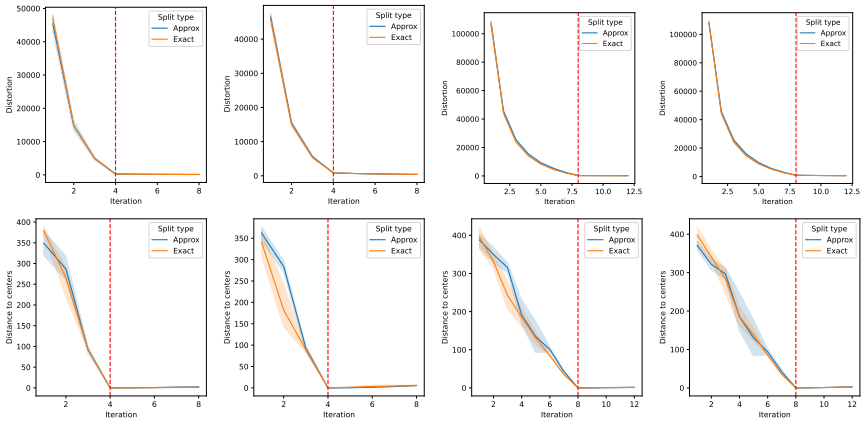
Table 2: Error (Kendall τ distance to the Borda local median) of Copeland and Schulze aggregation, across synthetic Mallows mixtures with varying n and number of components K.

Table 2 reports the error of Copeland and Schulze relative to the Borda median. Across all configurations, Copeland and Schulze yield essentially indistinguishable errors, with differences on the order of 0.4 to 4. Both errors remain very small relative to the maximal possible distance $n(n-1)/2$. Out of the 1200 experiments carried out, Copeland and Borda estimate the same ranking 778 times and 780 for Borda and Schulze. The reason is that the local median problem in COAST is solved only on low-variability cells since $\widehat{V}'_N(\mathcal{C}) < \epsilon$ triggers termination of COAST. In this regime, the three aggregation methods provide an accurate estimator of the Kemeny median.

B.4 Further experiments on the COAST

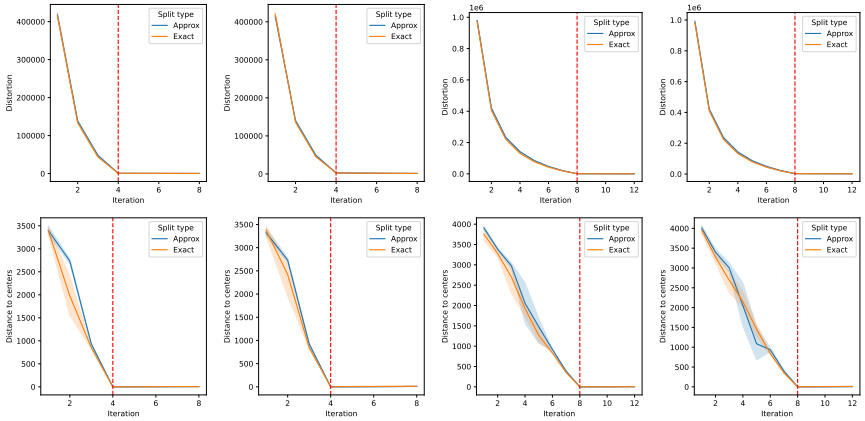
The experimental framework of the paper is detailed in this Appendix, where we consider a broader range of configurations and models, including the Plackett–Luce and Mallows models, see setting in the captions of Figures 5, 6 and 7. Two splitting criteria are evaluated, and both yield results of comparable quality. Although the alternative criterion is not optimal in the sense that it does not guarantee the largest decrease of the distortion at each iteration, both procedures ultimately converge to a similar level of distortion.

In addition, we report Table 3 summarizing the computational times (in seconds) obtained on a standard personal computer. The theoretical differences in computational complexity are reflected in practice: time results are similar for small n and k; however, for $n = 150$ the exact splitting method requires more than 15 minutes of computation, whereas the alternative criterion completes in approximately 15 seconds.



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

Figure 5: Components are Mallows models of $n = 50$, with $k = 4$ in (a) and (b), $k = 8$ in (c) and (d). Also, (a,c) are more concentrated than (b,d), with scale parameters $\phi = 0.1$ and 0.3 respectively.



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

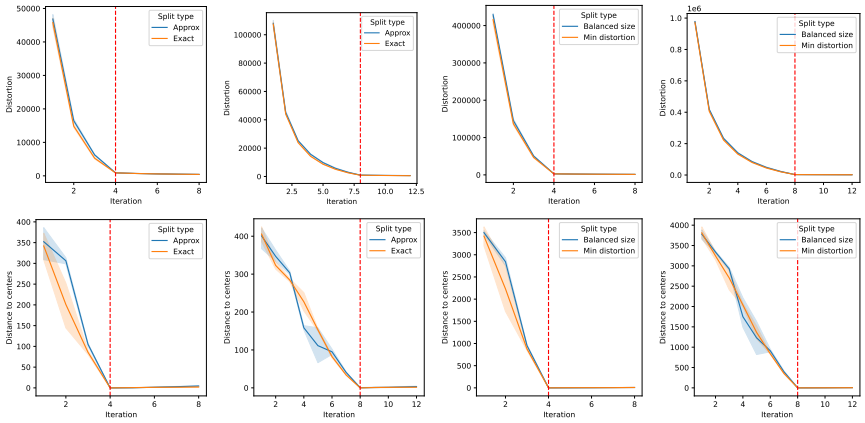
Figure 6: Components are Mallows models of $n = 150$, with $k = 4$ in (a) and (b), $k = 8$ in (c) and (d). Also, (a,c) are more concentrated than (b,d), with scale parameters $\phi = 0.1$ and 0.3 respectively.

B.4.1 Recovering the partitions in very noisy settings

In the main paper, we present the primary application of the COAST algorithm and study its ability to identify the components of a mixture through the evolution of the distortion across iterations, as well as through the distance between the recovered centers and the true centers of the underlying distribution. In the previous section, we consider the synthetic experiments in the case where the components are equally spread. Here, we extend this analysis to a noisier setting with heterogeneous scales across components.

We consider a sample drawn from a Mallows mixture (MM) distributions with $n = 20$ and $K = 4$ components and display the co-membership structure obtained after K iterations of the COAST algorithm. In Figure 8, the first row shows a heatmap of the pairwise distance matrix D, where D_{ij} denotes the Kendall's-τ distance between rankings σ_i and σ_j. The second row displays the cluster co-membership matrix $\tilde{D}$, where $\tilde{D}_{ij}$ is shown in green if and only if σ_i and σ_j belong to the same partition $\mathcal{C}$ after K iterations.

As can be observed, even under highly noisy conditions, the partitions corresponding to the original mixture components are almost perfectly recovered.



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

Figure 7: Components are Plackett-Luce of $n = 50$ (a,b) and $n = 150$ (c,d), with $k = 4$ (a,c), $k = 8$ in (b,d). All use exponential decrease of weights.

n	k	Balanced size	Min distortion
10	4	0.0826	0.0923
50	4	0.5197	3.5002
100	4	1.7686	48.4428
150	4	4.372	168.955
10	8	0.2424	0.3229
50	8	1.4242	13.9436
100	8	4.6943	161.861
150	8	15.022	1094.133

Table 3: Running times (in seconds)

B.5 Smoothing: depths are a good approximation for ranking distributions and CRD

In Figure 1(c) and related text in the main paper, we propose a smoothing technique for the ranking distribution within each component $\widetilde{P}_{\mathcal{C}}(\sigma)$ based on the local depth of the rankings associated with each partition $D_{P(\mathcal{C})}(\sigma)$. Since the expression of this local depth is trivial given that the local distribution can be approximated, $D_{P(\mathcal{C})}(\sigma) = \sum_{i<j} p_{\sigma^{-1}(i),\,\sigma^{-1}(j)}(\mathcal{C})$, the main challenge in this setting lies in the computation of the normalizing constant $Z_{\mathcal{C}}$. We give a closed form expression in the next proposition

Proposition 4. *The normalizing constant of the smoothed approximation is*

$$Z_{\mathcal{C}} = \sum_{i<j} p_{i,j}\big(1 - p'_{i,j}\big),$$

where p is $P(\mathcal{C})$ and $p' = U_{i,j}$ where U is given in Equation (36) *bellow.*

Proof. In order to give an analytical expression for $Z_{\mathcal{C}}$, we mimic the derivation of the expression for the variability of a distribution P written in the main paper as $V'_P = \sum_{i<j} p_{i,j}\big(1 - p_{i,j}\big)$: We recast this constant in terms of the variability induced by the uniform distribution on the partition $\mathcal{C}$. In particular, we write

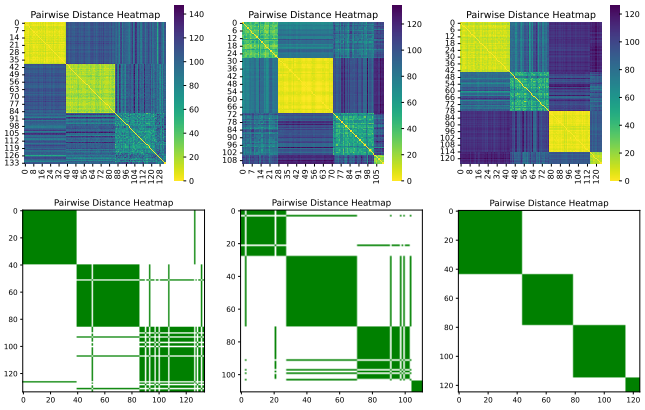
$$Z_{\mathcal{C}} = \frac{1}{2}\,\mathbb{E}\big[d(\Sigma,\Sigma')\big|\,\Sigma' \sim U_{\mathcal{C}}, \Sigma \sim P\big]\,. \tag{35}$$

This can be equivalently expressed as

$$Z_{\mathcal{C}} = \sum_{i<j} p_{i,j}\big(1 - p'_{i,j}\big),$$

where

$$p'_{i,j} = \mathbb{P}\big\{\Sigma'(i) < \Sigma'(j)\,\big|\,\Sigma \sim U'_{\mathcal{C}}\big\}.$$



(a) (b) (c)

Figure 8: Heatmap of the distances between pairs of rankings of noisy ranking models (Mallows model of random dispersion and $K = 4$ components, top row) and co-membership matrix after 4 iterations of the COST algorithm (bottom row).

The main difficulty in evaluating this expression lies in the computation of the probabilities $p'_{i,j}$. Closed-form expressions are trivial in several important cases, including $\mathcal{C} = S_n$, partially observed rankings [16], and the degenerate case $\mathcal{C} = \sigma$. In Equation (35) we are interested on $\mathcal{C} = U'_{\mathcal{C}}$, we derive a closed form expression for this in the next lines.

In the COAST algorithm with the alternative splitting criterion, the leaves of the tree consist on the partitions $\mathcal{C}$ that consist on a sequence of splits on pairs $(i, j) \in \mathcal{J}$ where each item i is only used once in the each $\mathcal{C}$. Let $U_{\mathcal{C}}$ denote the uniform distribution on the partition $\mathcal{C}$. Using basic enumerative combinatorics observations, it can be factorized as

$$U_{\mathcal{C}} = \prod_{(i,j)\in\mathcal{I}(\mathcal{C})} U_{i,j},$$

where $\mathcal{I}(\mathcal{C})$ denotes the set of unconstrained pairs induced by the partition $\mathcal{C}$. For each $(i, j) \in \mathcal{I}(\mathcal{C})$, the matrix $U_{i,j}$ is defined entrywise by

$$(U_{i,j})_{a,b} = \begin{cases} 1, & \text{if } a = i \text{ and } b = j, \\ \frac{1}{2}, & \text{if } a, b \neq i, j, \\ \frac{1}{3}, & \text{if } a = i \text{ and } b \neq j, \\ \frac{2}{3}, & \text{if } a \neq i \text{ and } b = j. \end{cases} \tag{36}$$

assuming $a < b$ and setting $(U_{i,j})_{b,a} = 1 - (U_{i,j})_{a,b}$. □

B.6 Visualization beyond ranking Depth: DD-plots on local vs. global depth

Visualizing distributions of rankings is particularly challenging, since standard tools such as histograms are not applicable. Originally introduced for multivariate data in [26] and [15], DD-plots were proposed as a visualization tool to illustrate variations in location and scale across different distributions. We extend this framework to the setting of CRD and general mixtures of rankings. Our goal is to visualize a CRD through their relative positions with respect to both the partition-wise (local) depth and the overall (global) depth, denoted by $D_{\hat{P}(\mathcal{C})}$ and $D_{\hat{P}_N}$, respectively.

Given a sample S drawn from a CDR model, we consider the associated empirical measures $\hat{P}_{\mathcal{C}}$, corresponding to a partition $\mathcal{C}$, and $\hat{P}_N$, associated with the entire sample. The DD-plot is constructed by representing each ranking $\sigma \in S$ as a point in the plane with coordinates

$$\big(D_{\hat{P}(\mathcal{C})}(\sigma),\, D_{\hat{P}_N}(\sigma)\big).$$

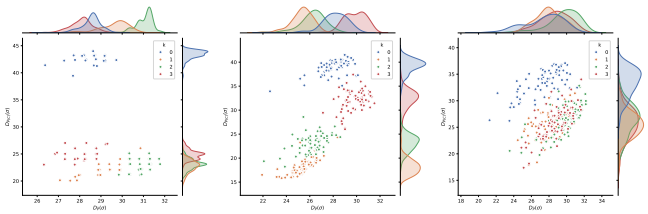
Points are color-coded according to the mixture component from which the ranking was generated. This representation enables a direct visual comparison between local and global depth behaviors,

thereby revealing structural features of the CRD, such as separation, overlap, or heterogeneity among components, that may not be apparent when relying on a single notion of depth.

In Figure 9, we illustrate the DD-plot obtained from samples drawn from a mixture of permutations of size $n = 10$ with $K = 4$ components, ranging from a highly peaked distribution (left panel) to a more uniform distribution (right panel). After K iterations of the COAST algorithm the partitioning structure recovered allows the evaluation of $(D_{\hat{P}(\mathcal{C})}(\sigma),\ D_{\hat{P}_N}(\sigma))$. The chosen local depth $D_{\hat{P}(\mathcal{C})}(\sigma)$ corresponds to cluster $k = 0$.

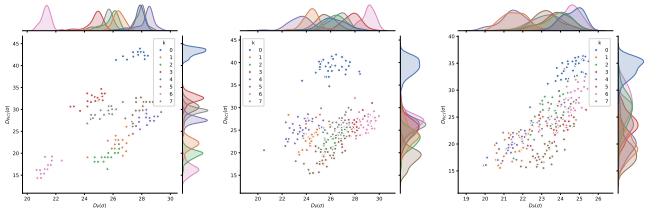
This graphical representation reveals a clear clustering structure: points tend to aggregate in specific regions of the DD-plot, reflecting the underlying clustering of the components in the data. This highlights the diagnostic power of the DD-plot for identifying structural heterogeneity in ranking samples. In particular, while observations are not clearly distinguishable when considering the global depth $D_{\hat{P}_N}$, they become well separated when local depth values $D_{\hat{P}(\mathcal{C})}$ are used.

The same conclusion are drawn for larger number of components, Figure 10, and Plackett-Luce model, Figure 11.



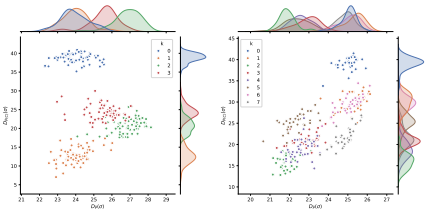
(a) (b) (c)

Figure 9: DD-plots on local vs. global depth: Mallows model $n = 10, k = 4$



(a) (b) (c)

Figure 10: DD-plots on local vs. global depth: Mallows model $n = 10, k = 8$



(a) (b)

Figure 11: DD-plots on local vs. global depth: Plackett-Luce for $n = 10$, $k = 4$ on (a) and $n = 10$, $k = 8$ on (b).

B.7 Anomaly detection

In section B.6 we show that the observations are not clearly distinguishable when considering the global depth $D_{\hat{P}_N}$, but they become well separated when local depth values $D_{\hat{P}(\mathcal{C})}$ are used. This idea motivates the use of local depths for the application of anomaly detection. In particular, we

consider 20 test samples from Mallows mixtures (MM), as well as Plackett-Luce mixtures, see setting in the caption of Figure 12.

We consider 4 inlier rankings and 4 outliers and we plot their local depths (y-axis) along iterations (x-axis) in Figure 12, clearly showing that the difference in local depth between inliers and outliers increases until the optimal number of partitions is reached.

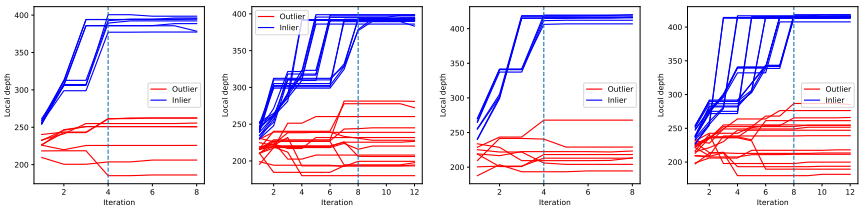


Figure 12: Results of $n = 30, \phi = .3$ and $K = 4$ (a) and $K = 8$ (b), Plackett-Luce mixtures (PL) with exponentially decreasing weights, $n = 30$ and $K = 4$ (c) and $K = 8$ (d).

B.8 Real-world data: Sushi dataset

We now evaluate the interpretability and performance of COAST on real-world data, specifically the Sushi dataset [21]. Figure 13(a) shows the evolution of the distortion of the COAST algorithm over 30 iterations under the two splitting criteria: the exact and the approximate criterion. The approximate criterion was run 10 times. After approximately 20 iterations, the distortion plateaus, indicating that most of the variability in the data has already been captured. Both criteria yield very similar distortion values, demonstrating that the approximation is effective.

Furthermore, the approximate criterion enables additional analysis through the expression $\widetilde{P}_{\mathcal{C}}(\sigma) = Z_{\mathcal{C}}^{-1} \sum_{i<j} p_{\sigma^{-1}(i),\sigma^{-1}(j)}(\mathcal{C})$, $\sigma \in \mathcal{C}$, with $Z_{\mathcal{C}}$ computable in $O(n^2 \cdot \text{depth}(\mathcal{C}))$ (Appendix B.5), as introduced in the experimental section of the main paper. In particular, the smoothing function can be interpreted probabilistically, allowing the use of a likelihood-based criterion, as shown in Figure 13(b) from which the same conclusion can be confirmed.

Figure 14 shows the partition structure of COAST under the exact splitting criterion. The results of the first iterations of the algorithm, using the deterministic criterion "Min distortion," are depicted as a tree. The first partition separates the population based on sushi types 4 and 3. This indicates that individuals who rank sushi 3 before sushi 4 form a distinct subgroup, and consequently, their rankings of the remaining sushi types differ from those who prefer sushi 4 to sushi 3. For each leave $\mathcal{C}$ in the tree from left to right, $P(\mathcal{C})$ is 0.13, 0.14, 0.12, 0.14, 0.12, 0.12, 0.1, 0.13, and their in-component variability V' is 8.88, 8.81, 8.92, 8.72, 9.0, 8.87, 8.81, 8.69.

The key takeaway is that this tool not only performs clustering but also serves as a valuable instrument for gaining insight into the underlying structure of the data and understanding the relationships between different elements in the ranking problem.

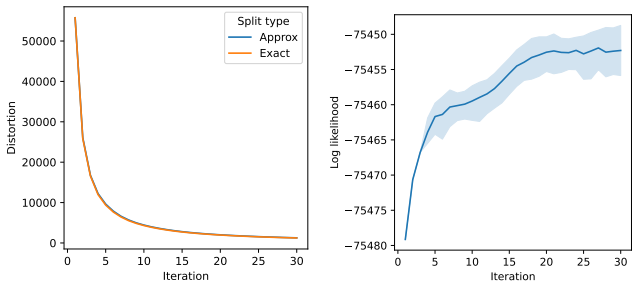


Figure 13: Distortion and likelihood on the Sushi dataset

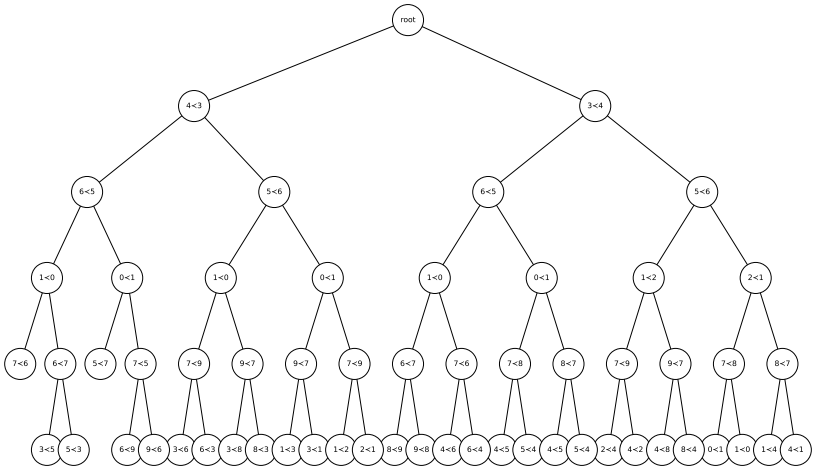


Figure 14: Tree structure recovered after 8 iterations of the COAST algorithm

B.9 LLM-benchmarking: TAC-08

Figure 15 gives the full tree of the TAC-08 dataset collected on the context of benchmarking LLMs, [5].

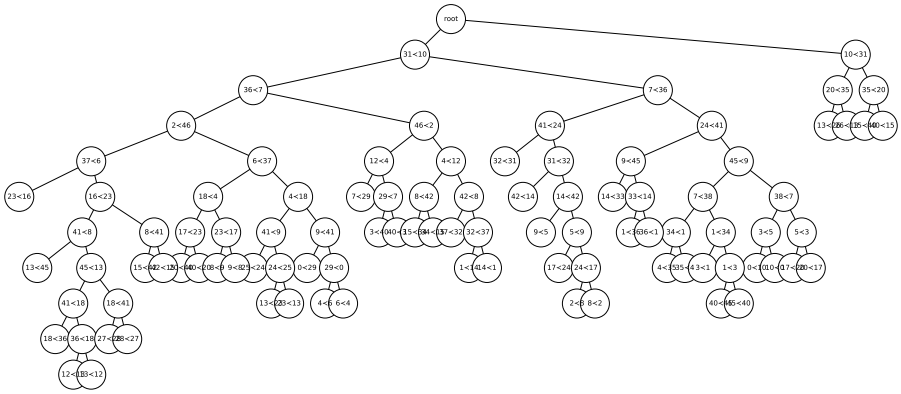


Figure 15: Tree structure recovered after 50 iterations of the COAST algorithm

C Limitations

In this section, we discuss the limitations and future directions of this paper. The results for large n are in Appendix B.4, which already reports experiments up to n = 150, Table 3. At that scale, the bottleneck is not the learning procedure itself but the sampling of synthetic data used to generate our test distributions, a numerical issue that can be addressed with dedicated sampling libraries (e.g., for Mallows/Plackett-Luce mixtures) rather than any limitation of COAST. The learning step is unaffected by this and retains the same computational complexity, namely $O(\#\mathcal{J}(\mathcal{C}))$ per split under the approximate splitting rule, as already discussed in the main paper (Remark 3).

For simplicity of exposition, the paper develops the methodology in the setting of complete rankings. The approach, however, extends naturally to partial rankings and to pairwise comparison data more generally. This is because the entire strategy is built upon the distortion criterion, which is computed from pairwise relations alone. Consequently, as long as the relevant pairwise relations $\Sigma(i) < \Sigma(j)$

can be observed, whether directly, as in pairwise comparison data, or as a byproduct of a partial ranking, the local pairwise probabilities $p_{i,j}(C)$, the distortion criterion $\mathcal{E}'_{(i,j)}(C)$, and hence the splitting rule underlying COAST, can all be computed exactly as in the complete ranking case. The recursive partitioning procedure and its theoretical guarantees therefore carry over directly to this more general setting.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made match theoretical and experimental results

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: the assumptions are clearly stated. Its tested in real and synthetic data. Computational efficienly is analyzed in theory and practice. For the limitations, alternatives are discussed. There is no privacy or fairness concern.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Assumptions, statements and proofs are clearly stated.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer:[Yes]

   Justification: The code and data to replicate all experiments is openly distributed.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code and data to replicate all experiments is openly distributed.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Replicability is ensured in the code, main paper and supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Experiments are repeated and averages and confidence intervals shown.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Experiments are done on a personal computer.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer:[Yes]

   Justification: no misalignment with the ethical code, no ethical concern.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: no societal impact of the work performed

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: the paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The authors are the owners.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer:[Yes]

    Justification: Drtaset, code and models are well documented.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: the core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.