Kantorovich distances between rankings with applications to rank aggregation

Stéphan Clémençon[1] and Jérémie Jakubowicz[1]

LTCI, Telecom Paristech (TSI) - UMR Institut Telecom/CNRS 5141
stephan.clemencon@telecom-paristech.fr
jeremie.jakubowicz@telecom-paristech.fr

Abstract. The goal of this paper is threefold. It first describes a novel way of measuring disagreement between rankings of a finite set $\mathcal{X}$ of $n \geq 1$ elements, that can be viewed as a (mass transportation) *Kantorovich metric*, once the collection rankings of $\mathcal{X}$ is embedded in the set $\mathcal{K}_n$ of $n \times n$ *doubly-stochastic* matrices. It also shows that such an embedding makes it possible to define a natural notion of *median*, that can be interpreted in a probabilistic fashion. In addition, from a computational perspective, the convexification induced by this approach makes median computation more tractable, in contrast to the standard metric-based method that generally yields NP-hard optimization problems. As an illustration, this novel methodology is applied to the issue of ranking aggregation, and is shown to compete with state of the art techniques.

1 Introduction

Formulated more than two centuries ago in the context of emerging social sciences and voting theories [Fis73], the problem of aggregating binary relations, (pre-) orders in particular, has recently received much attention in the machine-learning literature, see [HFCB08], [FKM$^+$03] or [MPPB07] for instance. Various modern applications sparked off the revival of interest in this issue, ranging from e-commerce to information retrieval through spam-fighting and database middleware. Indeed, in a wide variety of information systems now, input or output data take the form of an ordered list of items: search-engines, recommending systems, *etc.* Numerous tasks such as the design of meta-search engines, collaborative filtering, or combining results from multiple databases have motivated the development of new results in this domain, dedicated to three topics essentially: the extension of the notion of consensus among rankings [FKM$^+$06], the design of efficient algorithmic procedures for computing such median rankings, see [MM09] or [BFB09], and the building of probabilistic models on sets of rankings [LL03].

The present paper addresses all these aspects of the consensus problem, from an original angle. Its primary purpose is to show how the problem of measuring disagreement between rankings can be cast in terms of discrete *mass transportation problems*, by embedding the set of permutations in a convenient convex set of matrices. We prove that the continuum of metrics thus defined includes some

classical permutation metrics, such as the Hamming distance, the Spearman ρ distance or the Spearman footrule distance. From the perspective of rank aggregation, a novel (probabilistic) notion of median is next defined and related computational issues are tackled, taking advantage of the convexification step.

The paper is organized as follows. Notations are set out in Section 2, where most concepts involved in the subsequent analysis are introduced and an example motivating the present approach is also discussed. A novel way of measuring agreement between rankings is then proposed in Section 3, together with a definition of a probabilistic version of the notion of median ranking in Section 4. Results describing the computational complexity of the aggregation method proposed are stated in Section 5, while an illustrative application is presented in Section 6 Technical details are deferred to the Appendix.

2 Preliminary background

It is the purpose of this section to introduce the main concepts and definitions that shall be used throughout the paper.

2.1 First Definitions and Notation

We start off by recalling some definitions and setting out the notations needed in the subsequent analysis. Here and throughout, $\mathbb{I}\{\mathcal{E}\}$ denotes the indicator function of any event $\mathcal{E}$.

Rankings and matrix spaces. Let $n \geq 1$. We denote by $\mathfrak{S}_n$ the symmetric group of order n, *i.e.* the group of permutations of $\{1, \ldots, n\}$, and by $\mathcal{M}_n(\mathbb{R})$ the space of $n \times n$ matrices with real entries. Any permutation $\sigma \in \mathfrak{S}_n$ can be classically represented by the matrix

$$M^\sigma = (\mathbb{I}\{\sigma(i) = j\})_{1 \leq i,\, j \leq n}\,,$$

in $\mathcal{M}_n(\mathbb{R})$, whose entry $M^\sigma_{i,j}$ indicates whether rank j is assigned to the object indexed by i or not. The elements of the set $\Sigma_n = \{M_\sigma : \sigma \in \mathfrak{S}_n\}$ are called *permutation matrices*.

Medians. Given a collection $\mathit{\Pi} = \{\sigma_1, \ldots, \sigma_K\} \subset \mathfrak{S}_n$ of permutations (one commonly uses the term *profile* in social choice theory), the issue of summarizing the orders defined by $\mathit{\Pi}$'s elements, by a "consensual" (pre-) order, is called the *aggregation problem*. The so-termed *metric approach* is the most popular method for defining such a consensus. It assumes that a certain distance δ on the set $\mathfrak{S}_n$ is given. One calls a *median ranking* for the profile $\mathit{\Pi}$ with respect to a subset $\mathcal{R} \subset \mathfrak{S}_n$ any ranking $\sigma^* \in \mathcal{R}$ such that:

$$\sum_{k=1}^{K} d(\sigma^*, \sigma_k) = \min_{\sigma \in \mathcal{R}} \sum_{k=1}^{K} d(\sigma, \sigma_k). \tag{1}$$

The study of metrics on rankings has a long history, for instance one may refer to Chapter 11 in [DD09] for an excellent account of distances on permutations. The following distances, originally introduced in the context of nonparametric hypothesis testing, are among the most widely used.

- **The Kendall τ distance.** Counting the number of "discording pairs", it is given by: $\forall(\sigma_1, \sigma_2) \in \mathfrak{S}_n^2$,

$$d_\tau(\sigma_1, \sigma_2) = \sum_{1 \leq i < j \leq n} \mathbb{I}\{(\sigma_1(i) - \sigma_2(i)) \cdot (\sigma_1(j) - \sigma_2(j)) < 0\}.$$

- **The Spearman ρ distance.** It corresponds to the l_2-metric: $\forall(\sigma_1, \sigma_2) \in \mathfrak{S}_n^2$,

$$d_2(\sigma_1, \sigma_2) = \left(\sum_{i=1}^{n} (\sigma_1(i) - \sigma_2(i))^2 \right)^{1/2}.$$

- **The Spearman footrule distance.** This is actually the l_1-distance between rank vectors: $\forall(\sigma_1, \sigma_2) \in \mathfrak{S}_n^2$,

$$d_1(\sigma_1, \sigma_2) = \sum_{i=1}^{n} |\sigma_1(i) - \sigma_2(i)| .$$

- **The Hamming distance.** This is the l_0-distance between rank vectors: $\forall(\sigma_1, \sigma_2) \in \mathfrak{S}_n^2$,

$$d_0(\sigma_1, \sigma_2) = \sum_{i=1}^{n} \mathbb{I}\{\sigma_1(i) \neq \sigma_2(i)\} .$$

Many other distances could be considered, such as the Cayley/Kemeny distance [Kem59], or so-termed *word metrics* more generally [How00]. The major barrier to practical implementation of this approach lies in the fact that it generally leads to NP-hard problems, see [Hud08] or [Wak98]. Notice in addition that uniqueness of the median is not guaranteed in general. One may easily check for instance that, considering the Kendall τ distance, any permutation $\sigma \in \mathfrak{S}_n$ is a median with respect to the set $\mathfrak{S}_n$ (see also the example given below).

Remark 1. (The ordinal approach) Metric-based techniques are by no means the sole approach to rank aggregation. The so-termed "ordinal approach" includes a wide variety of techniques for combining rankings or, more generally, binary relations. They return to the famous "Arrow's voting paradox" and consist of a series of duels (i.e. pairwise comparisons) as in Condorcet's methods or successive tournaments as in the celebrated proportional voting Hare system. Special attentions has recently been paid to such techniques in the context of preference learning ("Ranking by Pairwise Comparison" methods); see [HFCB08] for instance.

2.2 A simple example

The following example shows that, beyond the computational difficulties above mentioned, the metric-based approach may have important drawbacks. Let us regard the problem of aggregating/summarizing the permutations described by the rank vectors (1, 2, 3) and (3, 2, 1) in $\mathfrak{S}_3$ for instance. Considering Kendall τ medians with respect to $\mathfrak{S}_3$ is clearly not informative, any permutation except those two permutations being a median. Looking at the hexagon in Fig. 1, providing a natural representation of $\mathfrak{S}_3$ (adjacent vertices are at Kendall τ distance one from each other), one inevitably longs to define the median in the middle of the line segment connecting the opposite vertices. In other terms, the major drawback of the aforementioned metric-based approach does not lie in the metric considered itself, but rather in the fact that the search for a "barycenter" is restricted to the "curve-shaped" set $\mathfrak{S}_n$. The view developed subsequently provides a rigorous meaning to a definition of a median in the interior of the hexagon. We are going to incorporate some uncertainty/fuzziness to the notion of median ranks by enlarging/convexifying the original ensemble $\mathfrak{S}_n$, and next define well adapted metrics on the larger space thus obtained.

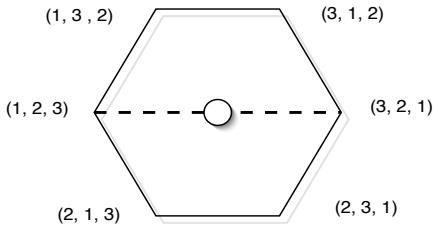

Fig. 1. Representation of the symmetric group $\mathfrak{S}_3$ as a regular hexagon.

2.3 Convexification/randomization

For clarity, we first recall the following definition.

Definition 1. (Double stochasticity) *A matrix $A = (a_{i,j}) \in \mathcal{M}_n(\mathbb{R})$ with nonnegative entries is said to be doubly stochastic if and only if*

$$\forall i \in \{1, \ldots, n\}, \ \sum_{j=1} a_{i,j} = \sum_{j=1} a_{j,i} = 1.$$

The set $\mathcal{K}_n$ of such doubly stochastic matrices is a convex subset of $\mathcal{M}_n(\mathbb{R})$.

Permutation matrices are special cases of doubly stochastic matrices. For clarity, let us recall the following celebrated result (see, for instance, [HJ85, p.539]).

Theorem 1 (Birkhoff–Von Neumann). *The set $\mathcal{K}_n$ is the convex hull of the set of permutation matrices Σ_n:*

$$\mathcal{K}_n = \mathrm{conv}\,(\Sigma_n)\,.$$

In addition, Σ_n corresponds to the set of $\mathcal{K}_n$'s extremal points.

Identifying $\mathfrak{S}_n$ with Σ_n, its embedding in $\mathcal{K}_n$ is a natural way of "convexifying" the rank aggregation problem. From the ranking perspective, the entry $a_{i,j}$ of a doubly stochastic matrix A can be interpreted as a marginal probability that rank j be assigned to the object No. i. Two standard ways of randomly generating a ranking from such a matrix are reviewed in subsections 4.1 and 4.2.

3 Kantorovich distances

We now introduce a general framework for measuring dissimilarity between rankings, following in the footsteps of the so-termed *mass transportation* approach to defining metrics between probability measures [Rac91].

3.1 Definitions and properties

We suppose now that we are given a certain (nonnegative) *cost function*, that is to say a mapping $c : \{1,\ \ldots,\ n\}^2 \times \{1,\ \ldots,\ n\}^2 \to \mathbb{R}_+ \cup \{+\infty\}$, $c((i,j),(k,l))$ representing the cost of moving one mass unit from (i,j) to (k,l). The technical conditions listed below shall be required in the subsequent analysis.

(*i*) (Diagonal terms) For all (i,j) in $\{1,\ \ldots,\ n\}^2$,

$$c\,((i,j),(i,j)) = 0.$$

(*ii*) (Symmetry) For all (i,j), (k,l) in $\{1,\ \ldots,\ n\}^2$,

$$c\,((i,j),(k,l)) = c\,((k,l),(i,j))\,.$$

(*iii*) (Triangular inequality) The cost function c on $\{1,\ \ldots,\ n\}^2$ fulfills the condition: for all (i,j), (k,l) and (s,t) in $\{1,\ \ldots,\ n\}^2$,

$$c((i,j),(k,l)) \leq c((i,j),(s,t)) + c((s,t),(k,l)).$$

(*iv*) (Non diagonal terms) For all $(i,j) \neq (k,l)$ in $\{1,\ \ldots,\ n\}^2$,

$$c\,((i,j),(k,l)) > 0.$$

Remark 2. Condition (*iii*) guarantees that the cost function c satisfies the (stronger in appearance) *reduction property*, meaning that

$$c((i,j),(k,l)) = \inf_{h \geq 1} c^h((i,j),(k,l)),$$

where, denoting by $\mathcal{P}_m((i,j),(k,l))$ the set of all paths of length m, $\{(u_m, v_m) : m = 0, \ldots, \}$, connecting (i,j) to (k,l), *i.e.* such that $(u_0, v_0) = (i,j)$ and $(u_{h+1}, v_{h+1}) = (k,l)$, we set: $\forall h \geq 1$,

$$c^h((i,j),(k,l)) = \inf \left\{ \sum_{m=1}^{h+1} c((u_{m-1}, v_{m-1})(u_m, v_m)) : \ (u,v) \in \mathcal{P}_m((i,j),(k,l)) \right\}.$$

Roughly, reduction amounts to state that no mass movement should be cheaper whenever performed in several steps rather than in one step.

Equipped with the notion of (symmetric and reduced) cost function, we may now define the concept of Kantorovich pseudo-metric on $\mathcal{K}_n$.

Proposition 1. (Mass transportation distance) *Let c be a cost function on $\{1, \ldots, n\}^2$ fulfilling conditions $(i) - (iii)$, $A = (a_{i,j})$ and $A' = (a'_{i,j})$ two elements of $\mathcal{K}_n^2$. If one defines the Kantorovich optimal transportation cost related to cost function c and real value $p \geq 1$ by:*

$$d_{c,p}(A, A') = \min_{\Phi \in \mathcal{M}(A,A')} \mu_{c,p}^{1/p}(\Phi), \tag{2}$$

with

$$\mu_{c,p}(\Phi) = \sum_{\substack{(i,j) \in \{1, \ldots, n\}^2 \\ (k,l) \in \{1, \ldots, n\}^2}} c\left((i,j),(k,l)\right)^p \Phi\left((k,l),(i,j)\right),$$

and where the set $\mathcal{M}(A, A')$ denotes the collection of mappings ("transportation plans") $\Phi : \{1, \ldots, n\}^2 \times \{1, \ldots, n\}^2 \to [0,1]$ such that: $\forall (i,j) \in \{1, \ldots, n\}^2$,

$$\sum_{(k,l) \in \{1, \ldots, n\}^2} \Phi((i,j),(k,l)) = a_{i,j} \ \text{ and } \sum_{(k,l) \in \{1, \ldots, n\}^2} \Phi((k,l),(i,j)) = a'_{i,j}. \tag{3}$$

Then $d_{c,p}$ is a pseudo-metric on $\mathcal{K}_n$: it satisfies the separability, symmetry and triangular inequality properties, but might fail to be always finite.

Obviously, $d_{c,p}$ being a pseudo-metric on $\mathcal{K}_n$, it is a pseudo-metric on $\mathfrak{S}_n$ (identified as Σ_n) as well, we set in this case $d_{c,p}(M^{\sigma_1}, M^{\sigma_2}) = d_{c,p}(\sigma_1, \sigma_2)$, with a slight abuse of notation. Before showing several important examples of such pseudo-metrics, a few remarks are in order.

Remark 3. (Normalization) For the sake of simplicity, this definition above uses a slightly different convention than in the classical mass transportation setting (see, for instance, [RR98,Vil09]). Indeed A and A' do not define probability measures on $\{1, \ldots, n\}^2$, their mass with respect to the counting measure being equal to n (it would simply suffice to divide the latter by n for leading back to the usual setup).

Remark 4. (MONGE VS. KANTOROVICH) When the search for transportation plans with minimum cost is restricted to plans Φ taking their values in $\{0,1\}$ (one does not try to divide the mass described by the entries of the initial matrix to transport it, assigning new locations to the original entries being sufficient in this situation), the problem is said of *Monge*'s type. We point out that, even when both the initial and final distributions of mass are described by permutation matrices M^{σ_1} and M^{σ_2}, OTP's for the Kantorovich problem are not Monge transportation plans in general. Indeed, consider for instance the simple case where $n = 2$ and the cost is constant, equal to some fixed scalar $\gamma > 0$, except on the diagonal $\{(i,j) = (k,l)\}$ where it is 0 (as required by condition (i)). It is easy to see that the optimal transportation cost between $\iota = (1\ 2)$ and $\tau = (2\ 1)$ is $2^{1/p}\gamma$ (identifying Σ_n with $\mathfrak{S}_n$). Additionally, observe that every transportation plan achieves this cost and only two of them are of Monge type.

Remark 5. (ON UNIQUENESS) It should be pointed up that optimal transportation plans are not necessarily unique (refer for instance to the example mentioned in the preceding remark).

Remark 6. (RANKING STABILITY) In the same way as Wasserstein-Kantorovich probability metrics turned to be quite adapted to the study of stability of stochastic models such as queuing systems (see [Rac91]), the optimal properties of distances $d_{c,p}(.,.)$ make them very useful for investigating the stability of ranking algorithms/models in a proper way. By a ranking algorithm, we mean here a mapping $\sigma : \mathcal{D} \mapsto \sigma_{\mathcal{D}}$ that assigns a permutation $\sigma_{\mathcal{D}} \in \mathfrak{S}_n$ to any training data sample $\mathcal{D}_N$ of size $N \geq 1$, allowing for ranking n objects, indexed by $i = 1, \ldots, n$. The nature of the sample may vary depending on the application considered (collaborative filtering or nonparametric scoring for instance), it can be made of preferences, rankings, binary data, *etc.* (see [CV09] and the references therein for instance). The definition below does not require to specify the nature of the training data however. Given a cost function c on $\{1, \ldots, n\}^2$, we define the instability measure as:

$$\mathbf{Instab}_N(\sigma) = \mathbb{E}_{\mathcal{D}_N, \mathcal{D}'_N}\left[d_{c,p}(\sigma_{\mathcal{D}_N}, \sigma_{\mathcal{D}'_N})\right],$$

where $\mathcal{D}'_N$ denotes an independent copy of the sample and $\mathbb{E}_{\mathcal{D}_N, \mathcal{D}'_N}[.]$ denotes the expectation taken with respect to $(\mathcal{D}, \mathcal{D}')$. We mention incidentally that such an instability measure can be estimated through a standard resampling scheme. By drawing data with replacement among the original sample, one may get $B \geq 1$ bootstrap replicates $\mathcal{D}^{*(1)}, \ldots, \mathcal{D}^{*(B)}$ of the sample $\mathcal{D}_N$. A bootstrap estimate of $\mathbf{Instab}_N(\sigma)$ is then given by:

$$\widehat{\mathbf{Instab}}_N(\sigma) = \frac{2}{B(B-1)} \sum_{1 \leq b < b' \leq B} d_{c,p}(\sigma_{\mathcal{D}^{*(b)}}, \sigma_{\mathcal{D}^{*(b')}}).$$

3.2 Examples

As proof of relevance of the approach embraced in this paper, we now show that some widely used metrics for measuring disagreement between rankings on

$\{1, \ldots, n\}$, can be viewed as restrictions to Σ_n of a Kantorovich distance (for an adequate choice of the cost function). A few important examples are listed below.

1. **Hamming distance.** It corresponds to the cost function

$$c_{\mathcal{H}}\left((i,j),(k,l)\right) = \begin{cases} 0 & \text{if } i=k, j=l \\ 1 & \text{if } i=k, j\neq l \\ +\infty & \text{otherwise} \end{cases},$$

   with $p=1$ in the sense that: $\forall(\sigma_1,\sigma_2)\in\mathfrak{S}_n^2$, $\delta_{\mathcal{H}}(\sigma_1,\sigma_2) = d_{c_{\mathcal{H}},1}(\sigma_1,\sigma_2)$.
2. **Spearman footrule distance.** It corresponds to the cost function

$$c\left((i,j),(k,l)\right) = \begin{cases} |j-l| & \text{if } i=k \\ +\infty & \text{otherwise} \end{cases},$$

   with $p=1$.
3. **Spearman ρ metric.** It corresponds to the same cost function as above, except that $p=2$ here.

The assertions above are easy to prove, $\mathcal{M}(M^{\sigma_1}, M^{\sigma_2})$ containing, in each case, a single element only.

Beyond the fact they can be seen as extensions of numerous permutation distances, the major advantage of the collection of Kantorovich pseudo-metrics lies in the considerable flexibility it provides for measuring disagreement between rankings. By choosing the cost properly, one may attach much more importance to the top ranks than to the others for instance, which makes sense in various rank aggregation tasks.

However, we suspect that not every classical distance on permutations can be recovered as a Kantorovich distance. Let us first introduce the following notions.

Definition 2. *A function f defined on $\mathcal{K}_n^2$ is said 'right-invariant' (respectively, 'left-invariant'), when: $\forall\sigma\in\mathfrak{S}_n$, $\forall A\in\mathcal{K}_n$,*

$$f(A\cdot\sigma, A'\cdot\sigma) = f(A,A') \quad (\textit{respectively,} \quad f(\sigma\cdot A, \sigma\cdot A') = f(A,A'))$$

where $A\cdot\sigma = (A_{i,\sigma(j)})$ (respectively, $\sigma\cdot A = (A_{\sigma(i),j})$). The function f is said bi-invariant when it is right-invariant and left-invariant both at the same time.

Equipped with these definitions, we state the following result, relating invariance properties of a cost function to those of the related pseudo-metric. Owing to space limitations, the proof is omitted and left to the reader.

Proposition 2. *Let c be a cost function fulfilling conditions $(i)-(iv)$ and n denote a large enough integer. The pseudo-metric d_c is bi-invariant if and only if the function c is bi-invariant when $\mathfrak{S}_n$ acts on $\{1,\ldots,n\}^2$ on the right by $(i,j)\cdot\sigma = (i,\sigma(j))$ (respectively, on the left by $\sigma\cdot(i,j) = (\sigma(i),j)$).*

Corollary 1. *There exists a nonnegative integer n_0, such that for all $n \geq n_0$, the Cayley distance between two permutations in $\mathfrak{S}_n$ (defined as the minimum number of transpositions to be composed with one of them to turn it into the other) is not the restriction to Σ_n of any Kantorovich distance on $\mathcal{K}_n$.*

4 From medians in $\mathcal{K}_n$ to median rankings

Now the concept of Kantorovich metric between rankings has been introduced, our main goal is to use it in order to define and compute *medians*, summarizing a profile of rankings, in $\mathcal{K}_n$ first, and in $\mathfrak{S}_n$ next.

Definition 3. *Let $n \geq 1$ and let $A_1, \ldots, A_N$ be $N \geq 1$ elements of $\mathcal{K}_n$. Any matrix $A^* \in \mathcal{K}_n$ such that*

$$\sum_{m=1}^{N} d_{c,p}(A^*, A_m) = \inf_{A \in \mathcal{K}_n} \sum_{m=1}^{N} d_{c,p}(A, A_m) \tag{4}$$

is called a median matrix for the profile $\{A_1, \ldots, A_N\}$.

Remark 7. (On existence) We point out that medians, in the sense of Definition 3, always exist. Indeed, for any $(A_1, \ldots, A_N) \in \mathcal{K}_n^N$, $N \geq 1$, the mapping $A \in \mathcal{M}_n(\mathbb{R}) \mapsto \sum_{m=1}^{N} d_c(A, A_m)$ is continuous, the infimum over the compact set $\mathcal{K}_n$ being thus achieved. In contrast, regarding uniqueness, we underline that in general several medians may exist.

By means of this definition, given a profile $(A_m)_{1 \leq m \leq N}$ in $\mathfrak{S}_n$, we end up with a summary median matrix A^* in $\mathcal{K}_n$, which, in general, does not lie in Σ_n. This is the convexification step. When trying to summarize the statistical properties of the profile, all the useful information is encoded in this 'central matrix'. However, when willing to perform certain specific tasks related to rank aggregation, it is desirable to recover a ranking, not a matrix in $\mathcal{K}_n$. We review below two popular approaches for building a 'median ranking' based on a median matrix.

4.1 The Mallows model

Let $A^* = (a^*_{i,j}) \in \mathcal{K}_n$ be fixed. A flexible approach is to generate a ranking σ, of which permutation matrix M^σ is 'close to A^*' (in the sense of a Kantorovich pseudo-metric $d_{c,p}$), consists in drawing at random an element from $\mathfrak{S}_n$ so that the smaller the distance $d_{c,p}(A^*, M^\sigma)$, the larger the probability of occurence. This is exactly the purpose of the celebrated Mallows model [Mal57,Dia89], that consists, in our context, to consider the distribution given by: $\forall \sigma \in \mathfrak{S}_n$,

$$\mathbb{P}\{\sigma\} = \frac{1}{Z} \exp\left(-\theta d_{c,p}(M^\sigma, A^*)\right),$$

where Z is a normalization constant and θ is a positive parameter. When $1/\theta$ is small compared to the distance $d_{c,p}(A^*, \Sigma_n)$ only the nearest profiles are given a chance to be drawn, when it is large more distant profile are likely to appear. The main drawback of this model lies in its huge computational complexity, in $O(n!)$ namely.

When $1/\theta$ tends to 0, the Mallows model degenerates towards the uniform distribution on the set $\{\sigma \in \mathfrak{S}_n : d_{c,p}(A^*, M^\sigma) = d_{c,p}(A^*, \Sigma_n)\}$, where we set $d_{c,p}(A^*, \Sigma_n) = \min_{M \in \Sigma_n} d_{c,p}(A^*, M)$ by definition.

4.2 Variants of the Luce model

Probabilistic approach. A more lightweight approach consists in reinterpreting the median matrix entries as scores and relying then on an adaptation of the *Luce model*, see [Luc59,Pla75]. More precisely, we start off with choosing randomly the object i_1 ranked first, by drawing according to the distribution on $\{1, \ldots, n\}$ defined by the column $(a^*_{i,1})_{1 \leq i \leq n}$ of A^*, then we generate the object ranked second among the remaining ones $\{1, \ldots, n\} \setminus \{i_1\}$ according to the distribution $(a^*_{i,1} + a^*_{i,2})/(2 - a^*_{i_1,1})$ for $i \neq i_1$, and so on and so forth. Formally, the distribution of the ranking σ drawn from this model can be written as,

$$\mathbb{P}\left\{\sigma^{-1}(1, \ldots, n) = (i_1, \ldots, i_n)\right\} = \prod_{k=1}^{n} f_k(i_k;\, i_{k-1}, \ldots, i_1),$$

for all permutation $(i_1, \ldots, i_n)$ of $(1, \ldots, n)$, where

$$f_k(i_k;\, i_{k-1}, \ldots, i_1) = \frac{\sum_{j=1}^{k} a^*_{i_k,j}}{\sum_{i \notin \{i_1, \ldots, i_{k-1}\}} \sum_{j=1}^{k} a^*_{i,j}}.$$

Of course, in a dual fashion, we could draw at random the rank assigned to the object 1, according to the distribution $(a^*_{1,j})_{1 \leq j \leq n}$, *etc.*

Greedy approach. A greedy version of the approach described above can also be considered. Precisely, it consists in using the model

$$\mathbb{P}\left\{\sigma^{-1}(1, \ldots, n) = (i_1, \ldots, i_n)\right\} = \prod_{k=1}^{n} g_k(i_k;\, i_{k-1}, \ldots, i_1)$$

where

$$g_k(i_k;\, i_{k-1}, \ldots, i_1) = \frac{\mathbb{I}\left\{\sum_{j=1}^{k} a^*_{i_k,j} = \max_{i \notin \{i_1, \ldots, i_{k-1}\}} \sum_{j=1}^{k} a^*_{i,j}\right\}}{\sum_{i' \notin \{i_1, \ldots, i_{k-1}\}} \mathbb{I}\left\{\sum_{j=1}^{k} a^*_{i',j} = \max_{i \notin \{i_1, \ldots, i_{k-1}\}} \sum_{j=1}^{k} a^*_{i,j}\right\}}$$

The computational cost of both procedures is linear in the size of A^* (*i.e.* $O(n^2)$), insofar as the sums $s_{i,k} = \sum_{j=1}^{k} a_{i,j}$ are precomputed.

5 Computational aspects

It is the purpose of this section to investigate the computional complexity of the key ingredients of the methodology proposed precedingly: distances and medians.

5.1 Computing Kantorovich distances in $\mathcal{K}_n$

To compute $d_{c,p}$ with no other assumptions on the cost c than (i), (ii), (iii) and (iv), we make use of linear programming via an interior-point method [BGLS03], which solves the minimization problem (2) in weak polynomial time. If c only takes integer values, then the Edmonds-Karp algorithm is known to solve the problem in $O(n^2(m + n^2 \log n))$ time, m denoting the number of entries $((i, j), (k, l))$ where the cost is finite (there are at most n^4 such entries), see [KV00]. However, for some specific choices of c, the computational cost may reduce to $O(n)$, as it is the case for the Hamming and Spearman distances restricted to Σ_n, as already seen in Section 3.

5.2 Computing median matrices

In absence of any further assumptions on c (except conditions $(i)-(iv)$) and p, we resort to general non-linear programming method for solving the minimization problem (4), with no guarantee of convergence to an optimal solution. For specific cost functions and configurations, one can compute the median matrix efficiently, in polynomial time, as claimed in the next proposition.

Proposition 3. *Let $N \in \mathbb{N}^*$ and σ_m are permutations of $\mathfrak{S}_n$ for $m \in 1 \dots n$. Assume also that $\forall i \in 1 \dots n$, $m \neq m'$ implies $\sigma_m(i) \neq \sigma_{m'}(i)$. Then, equipped with the cost $c_{\mathcal{H}}$, $\bar{M} = 1/n \sum_{m=1}^{N} M^{\sigma_m}$ is a median matrix of $(M^{\sigma_m})_{m \in 1 \dots N}$ for $d_{c_{\mathcal{H}}}$. It can be computed with complexity $O(n^2 N \wedge nN \log(nN))$.*

6 Applications to rank aggregation

The advantages of the approach to the 'consensus issue' proposed in the preceding sections are now illustrated on numerical experiments related to the so-termed 'rank aggregation task' for meta-search engines in the context of Information Retrieval (IR) applications.

6.1 Datasets

Our experimental study is based on the LETOR database ([LXQ+07]). It is a public data repository, created for evaluating ranking algorithms. There are two 'rank aggregation' datasets in LETOR: `MQ2007-agg` and `MQ2008-agg`. Both datasets are of the same format, but differ in size and in number of ranks to be aggregated. A query q is submitted to N search engines; each engine outputs a list of pairs (document, score). Each line of the database is of the form:

`Relevance  QueryId  ScoreEngine1 ...  ScoreEngineN  DocId`

where *Relevance* is an integer between 0 (irrelevant) and 2 (highly relevant). With our previous notations, n changes at each query, we denote it by n_q, whereas N is constant within the dataset (21 for `MQ2007-agg`, 25 for `MQ2008-agg`). Table 1 contains some basic statistics about the data: the number of queries, engines, the average number of documents by query and the max number of documents by query. The goal is to find, for each query in the dataset, the best possible ranking, ranking accuracy being assessed by the means of the Normalized Discounted Cumulative Gain (NDCG) measure ([JK02]). Recall that the DCG, up to rank r, is defined by

$$\mathrm{DCG_r}(\sigma) = \sum_{\mathrm{i}=1}^{\mathrm{r}} \frac{2^{\mathrm{rel}\left(\sigma^{-1}(\mathrm{i})\right)} - 1}{\log_2(1+\mathrm{i})},$$

where rel denotes the relevance of document i for a given query, and σ the sought aggregated ranking. Now $\mathrm{NDCG_r}$ is the same quantity normalized so that we have $\mathrm{NDCG_r}(\sigma) = 1$ for the best ranking.

LETOR provides a benchmark with the `BordaCount` method ([AM01]).

dataset	MQ2007-agg	MQ2008-agg
#queries	1692	784
#engines	21	25
#avgdocs	41.1	19.4
#maxdocs	147	121

Table 1. Datasets statistics

6.2 Implementation Details

The (huge) size of the datasets considered lead us to rule out general cost functions and the use of the Mallows model. Instead, we chose the Hamming cost $c_{\mathcal{H}}$ and used Proposition 3 for median computations (when the hypothesis of disjoint supports of Proposition 3 is not satisfied, the computed 'median' solely consists of an approximation of the optimum). We then tested and compared the models of section 4.2 (except the Mallows model, too demanding in regards to the dataset size) to extract a ranking from the computed median matrix. For the degenerate Mallows model, we used the Hungarian method (Kuhn-Munkres algorithm) [Mun57]. It has complexity $O(n_q^3)$, where n_q denotes the number of documents associated to query q.

6.3 Results

LETOR datasets are organized into training, validation and test set. However, since, like `BordaCount` our method is unsupervised, we used the whole dataset as a test set without restriction.

NDCG	@1	@2	@3	@4	@5	@6	@7	@8	@9	@10
BordaCount	0.1902	0.2014	0.2081	0.2128	0.2188	0.2247	0.2312	0.2377	0.2444	0.2507
LUCE Greedy	0.1980	0.2058	0.2137	0.2229	0.2301	0.2379	0.2441	0.2505	0.2575	0.2648
LUCE Random	0.2275	0.2328	0.2406	0.2450	0.2515	0.2578	0.2623	0.2683	0.2745	0.2814
Mallows-∞	0.1920	0.2044	0.2100	0.2170	0.2226	0.2283	0.2346	0.2419	0.2471	0.2535

Table 2. Results comparisons on the `MQ2007-agg` dataset

NDCG	@1	@2	@3	@4	@5	@6	@7	@8	@9	@10
BordaCount	0.2368	0.2806	0.3080	0.3432	0.3713	0.3888	0.3992	0.3724	0.1643	0.1694
LUCE Greedy	0.2026	0.2563	0.3058	0.3426	0.3752	0.3936	0.4030	0.3749	0.1567	0.1644
LUCE Random	0.2188	0.2726	0.3005	0.3279	0.3498	0.3691	0.3833	0.3579	0.1456	0.1508
Mallows-∞	0.1937	0.2374	0.2787	0.3176	0.3487	0.3703	0.3841	0.3587	0.1459	0.1541

Table 3. Results comparisons on the `MQ2008-agg` dataset

The tables above show that, for both datasets, rank aggregation based on the Hamming-Kantorovich distance in $\mathcal{K}_n$ lead to competitive results, compared to the `BordaCount` procedure.

7 Conclusion

In this paper, we have provided a novel family of distances between rankings of a finite number of elements, which can be viewed as mass transportation distances, by the means of an embedding of the set of permutation matrices Σ_n in the set $\mathcal{K}_n$ of doubly-stochastic matrices. This convexification step is also shown to be a key ingredient for defining a new and flexible concept of median, reflecting a consensus among a finite number of rankings. Although the freedom in the choice of the cost function may lead to optimize a variety of tasks in the ranking context such as stability evaluation or ranking prediction, a simple application of this approach based on the Hamming cost yielded promising results, competing with those produced by the `BordaCount` method on LETOR benchmark datasets. Truth be told, this choice has been mainly motivated by computational convenience. Algorithmic issues concerning distance/median computation and properties of the median (Pareto efficiency, *etc.*), depending on the conditions fulfilled by the underlying cost, will be the subject of further research.

Appendix - Technical Proofs

Proof of Proposition 1

Observe first that, for any $(A, A') \in \mathcal{K}_n^2$, the quantity $d_c(A, A')$ is well-defined as a minimum, since the functional μ_c is linear (hence continuous) on $\{\Phi : \{1, \ldots, n\}^4 \to \mathbb{R}\}$ and thus continuous on $\mathcal{M}(A, A')$ as well, which space is

compact. Therefore a transportation plan Φ^* achieves the minimum (2). It is called an *optimal transportation plan* (OTP). The symmetry of d_c immediately results from the symmetry of the cost function c. The separability of d_c is an easy consequence of hypothesis (i) and (iv) for c. Let us finally prove the triangular inequality for d_c. Assume that A, A' and A'' are three given matrices in $\mathcal{K}_n$. Let us denote by Φ_1 an OTP from A to A'', and by Φ_2 an OTP from A'' to A'. From the *gluing lemma* [Vil09, p.23] there exists a map Φ_{132} from $\{1,\dots,n\}^3$ to $[0,1]$ such that $\forall (i,j,k,l) \in \{1,\dots,n\}^4$,

$$\sum_{(k,l)\in\{1,\dots,n\}^2} \Phi_{132}\left((i,j),(r,s),(k,l)\right) = \Phi_1\left((i,j),(r,s)\right)$$

and

$$\sum_{(i,j)\in\{1,\dots,n\}^2} \Phi_{132}\left((i,j),(r,s),(k,l)\right) = \Phi_1\left((r,s),(k,l)\right)$$

Let $\Phi\left((i,j),(k,l)\right) = \sum_{(r,s)\in\{1,\dots,n\}^2} \Phi_{132}\left((i,j),(r,s),(k,l)\right)$, then, triangular inequality of c and Minkowski inequality implies:

$$\begin{aligned} d_{c,p}(A,A') &\leq \mu_{c,p}^{1/p}(\Phi) \\ &= \left(\sum_{\substack{(i,j)\in\{1,\dots,n\}^2 \\ (k,l)\in\{1,\dots,n\}^2 \\ (r,s)\in\{1,\dots,n\}^2}} c\left((i,j),(k,l)\right)^p \Phi_{132}\left((k,l),(r,s),(i,j)\right) \right)^{1/p} \\ &\leq d_c(A,A'') + d_c(A'',A') \end{aligned}$$

Proof of Corollary 1

The Caley distance is bi-invariant. By virtue of Proposition 2, the cost c itself is bi-invariant. Now, under the action

$$(\sigma,\sigma') \in \mathfrak{S}_n \times \mathfrak{S}_n \mapsto \big(\big((i,j),(k,l)\big) \in \{1,\dots,n\}^4 \mapsto \big((\sigma(i),\sigma'(j)),(\sigma(k),\sigma'(l))\big)\big) \ ,$$

the set $\{1,\dots,n\}^4$ has 4 distinct orbits. The first orbit is the diagonal $D_n = \{\big((i,j),(i,j)\big) : (i,j) \in \{1,\dots,n\}^2\}$, the 3 other orbits are $H_n = \{\big((i,j),(i,l)\big) : j \neq l\}$, $V_n = \{\big((i,j),(k,j)\big)\}$ and $O_n = \{\big((i,j),(k,l)\big) : i \neq k, j \neq l\}$. On D_n, the cost c is necessarily zero due to condition (i). From $\mathfrak{S}_n^2$ invariance of c, we know that c takes a constant value h over H_n, (over V_n and over O_n respectively). Triangular inequality implies that $o \leq h + v$, but also that $h \leq 2o$ and $v \leq 2o$. We now distinguish two cases. Either $o < +\infty$ or else $o = +\infty$. If $o = +\infty$ then either v or h is infinite, in which case we have already seen that it corresponds to the Hamming distance on permutations (which is obviously bi-invariant too and different from the Cayley distance), since we rule out the case where $o = v = h = +\infty$. If $o < +\infty$, then h and v are also finite. Since both h and

v are finite, invariance also leads to $h = v$ using $(M^\sigma)^T = (M^\sigma)^{-1}$. Considering the distance between a transposition matrix and the identity, we deduce that $h = v = 1/2$. Now, considering the matrix corresponding to a length 3-cycle we deduce that $3o \leq 2$ (otherwise the cost of transportation from the identity to the cycle using only horizontal or vertical movements is larger than n whereas Cayley distance is equal to $n - 1$). But now there is a contradiction with the length n cycle whose transportation cost to the identity is less than $2n/3$ instead of $n - 1$ for the Cayley distance.

Proof of Proposition 3

It easy to see that the transportation distance between two vectors v and w in $\mathbb{R}^n$ induced by the cost $c(i, j) = \mathbb{I}\{i \neq j]\}$ is $\sum_{i=1}^{n} |w_i - v_i|/2$. Consider a fixed row index i in $1 \dots n$. By definition of $c_{\mathcal{H}}$, the transportation plan should not mix rows (otherwise the transportation cost would be infinite). The cost induced by row i between the matrix M^σ and M, where σ denotes any permutation of $\mathfrak{S}_n$ and M any matrix in $\mathcal{K}_n$, is then: $1 - M(i, \sigma(i))$. Hence, we have:

$$\sum_m d_{c_{\mathcal{H}}}(M, M^{\sigma_m}) \geq \sum_i \sum_m \left(1 - M(i, \sigma_m(i))\right).$$

Now, since M is doubly stochastic, we may write

$$\sum_i \sum_m \left(1 - M(i, \sigma_m(i))\right) = nN - \sum_i \sum_m M(i, \sigma_m(i)) \geq Nn - \sum_m \sum_i \bar{M}(i, \sigma_m(i))$$

. Since all the $\sigma_m(i)$'s are distinct (i being fixed), we have

$$\sum_m d_{c_{\mathcal{H}}}(M, M^{\sigma_m}) \geq \sum_m d_{c_{\mathcal{H}}}(\bar{M}, M^{\sigma_m}).$$

If $\log(nN) \leq n$, one may store all the matrices M^{σ_m} using a dictionary structure, where each lookup costs at most $\log(nN)$. Otherwise, one can simply sum up the matrices M^{σ_m}.

References

[AM01] J.A. Aslam and M. Montague. Models for metasearch. In SIGIR01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM Press, 2001.

[BFB09] M.S. Bansal and D. Fernandez-Baca. Computing distances between partial rankings. *Information Processing Letters*, 109:238–241, 2009.

[BGLS03] J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal. *Numerical optimization*. Universitext. Springer-Verlag, Berlin, 2003.

[CV09] S. Clémençon and N. Vayatis. Tree-based ranking methods. *IEEE Transactions on Information Theory*, 55(9):4316–4336, 2009.

[DD09] M.M. Deza and E. Deza. *Encyclopedia of Distances*. Springer, 2009.

[Dia89] P. Diaconis. A generalization of spectral analysis with application to ranked data. *The Annals of Statistics*, 17(3):949–979, 1989.

[Fis73] P. Fishburn. *The Theory of Social Choice*. University Press, Princeton, 1973.

[FKM+03] R. Fagin, R. Kumar, M. Mahdian, D. Sivakumar, and E. Vee. Comparing and aggregating rankings with ties. In *Proceedings of the 12-th WWW conference*, pages 366–375, 2003.

[FKM+06] R. Fagin, R. Kumar, M. Mahdian, D. Sivakumar, and E. Vee. Comparing partial rankings. *SIAM J. Discrete Mathematics*, 20(3):628–648, 2006.

[HFCB08] E. Hüllermeier, J. Fürnkranz, W. Cheng, and K. Brinker. Label ranking by learning pairwise preferences. *Artificial Intelligence*, 172:1897–1917, 2008.

[HJ85] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

[How00] J. Howie. Hyperbolic groups. *In Groups and Applications, edited by V. Metaftsis, Ekdoseis Ziti, Thessaloniki*, pages 137–160, 2000.

[Hud08] O. Hudry. NP-hardness results for the aggregation of linear orders into median orders. *Ann. Oper. Res.*, 163:63–88, 2008.

[JK02] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):446, 2002.

[Kem59] J. G. Kemeny. Mathematics without numbers. *Daedalus*, 88:571–591, 1959.

[KV00] B. Korte and J. Vygen. *Combinatorial optimization*, volume 21 of *Algorithms and Combinatorics*. Springer-Verlag, Berlin, 2000.

[LL03] G. Lebanon and J. Lafferty. Conditional models on the ranking poset. In *Proceedings of NIPS'03*, 2003.

[Luc59] R. D. Luce. *Individual Choice Behavior*. Wiley, 1959.

[LXQ+07] T.Y. Liu, J. Xu, T. Qin, W. Xiong, and H. Li. Letor: Benchmark dataset for research on learning to rank for information retrieval. In *Proceedings of SIGIR 2007 Workshop on Learning to Rank for Information Retrieval*, pages 3–10, 2007.

[Mal57] C. L. Mallows. Non-null ranking models. i. *Biometrika*, 44(1-2):114–130, June 1957.

[MM09] B. Mandhani and M. Meila. Tractable search for learning exponential models of rankings. In *Proceedings of AISTATS, Vol. 5 of JMLR:W&CP 5*, 2009.

[MPPB07] M. Meila, K. Phadnis, A. Patterson, and J. Bilmes. Consensus ranking under the exponential model. In *Conference on Artificial Intelligence (UAI)*, pages 729–734, 2007.

[Mun57] J. Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38, 1957.

[Pla75] R. L. Plackett. The analysis of permutations. *Applied Statistics*, 2(24):193–202, 1975.

[Rac91] S.T. Rachev. *Probability Metrics and the Stability of Stochastic Models*. Wiley, 1991.

[RR98] S.T. Rachev and L. Rüschendorf. *Mass Transportation Problems. Volume I: Theory*. Springer, 1998.

[Vil09] C. Villani. *Optimal transport*, volume 338 of *Grundlehren der Mathematischen Wissenschaften*. Springer-Verlag, 2009.

[Wak98] Y. Wakabayashi. The complexity of computing medians of relations. *Resenhas*, 3(3):323–349, 1998.