MRA-based statistical learning from incomplete rankings

Eric Sibony, Stéphan Clémençon, Jérémie Jakubowicz

MRA-based Statistical Learning from Incomplete Rankings

Eric Sibony ERIC.SIBONY@TELECOM-PARISTECH.FR
Stéphan Clémençon STEPHAN.CLEMENCON@TELECOM-PARISTECH.FR
LTCI UMR No. 5141, Telecom ParisTech/CNRS, Institut Mines-Telecom, Paris, 75013, France

Jérémie Jakubowicz JEREMIE.JAKUBOWICZ@TELECOM-SUDPARIS.EDU
SAMOVAR UMR No. 5157, Telecom SudParis/CNRS, Institut Mines-Telecom, Evry, 91000, France

Abstract

Statistical analysis of rank data describing preferences over small and variable subsets of a potentially large ensemble of items $\{1, \ldots, n\}$ is a very challenging problem. It is motivated by a wide variety of modern applications, such as recommender systems or search engines. However, very few inference methods have been documented in the literature to learn a ranking model from such incomplete rank data. The goal of this paper is twofold: it develops a rigorous mathematical framework for the problem of learning a ranking model from incomplete rankings and introduces a novel general statistical method to address it. Based on an original concept of *multi-resolution analysis* (MRA) of incomplete rankings, it finely adapts to any observation setting, leading to a statistical accuracy and an algorithmic complexity that depend directly on the complexity of the observed data. Beyond theoretical guarantees, we also provide experimental results that show its statistical performance.

1. Introduction

Motivated by numerous modern applications ranging from the design of recommender systems to customer analytics through the optimization of search engines, the analysis of rank data has recently been the subject of much attention in the machine-learning literature. A "full ranking" on a set of $n \geq 1$ items $[\![n]\!] := \{1, \ldots, n\}$ is an ordering of the form $a_1 \succ \cdots \succ a_n$, where a_1 and a_n are respectively the items ranked first and last. It is usually described as the permutation σ on $[\![n]\!]$ that maps an item to its rank, *i.e.* such that $\sigma(a_i) = i$ for all $i \in [\![n]\!]$. The variability of a statistical population of full rankings is thus modeled by a probability distribution p on the symmetric group $\mathfrak{S}_n$, the set of all permutations on $[\![n]\!]$, called a *ranking model*. The statistical issue is then to learn this ranking model from the available observations, either through parametric modeling such as those proposed in the seminal contributions of (Mallows, 1957), (Luce, 1959; Plackett, 1975) or more recently in (Fligner & Verducci, 1986; Liqun, 2000; Lebanon & Lafferty, 2002; Meek & Meila, 2014), or through "nonparametric-like" approaches (*i.e.* without assuming an explicit structure for the distribution), such as kernel estimation (Lebanon & Mao, 2008; Sun et al., 2012), techniques based on sparsity concepts (Jagabathula & Shah, 2011), independence modeling (Huang & Guestrin, 2009) or inference based on harmonic analysis of the space of distributions on $\mathfrak{S}_n$ (Diaconis, 1988; Huang et al., 2009; Kondor & Barbosa, 2010; Irurozki et al., 2011; Kakarala, 2011; Kondor & Dempsey, 2012).

However, a major issue arises in practice from the fact that rank data seldom take the form of full rankings. Much more frequently the available rank data describe "limited" information, of the form of *partial* and/or *incomplete rankings*. Formally, they correspond to partial orders on $[\![n]\!]$ and can be naturally identified as subsets of permutations on $[\![n]\!]$, the subsets formed by their linear extensions. Let p be the probability distribution on $\mathfrak{S}_n$ of the random permutation Σ that models the variability of the preferences among the population of interest. The probability of a partial order represented by $S \subset \mathfrak{S}_n$ is then defined by

$$\mathbb{P}[\Sigma \in S] = \sum_{\sigma \in S} p(\sigma).$$

For instance, the probability that item i be ranked first is given by $\mathbb{P}[\Sigma(i) = 1] = \sum_{\sigma \in \mathfrak{S}_n,\, \sigma(i)=1} p(\sigma)$, and the probability that item i be preferred to item j is given by $\mathbb{P}[\Sigma(i) < \Sigma(j)] = \sum_{\sigma \in \mathfrak{S}_n,\, \sigma(i)<\sigma(j)} p(\sigma)$. These quantities correspond to *marginal probabilities* of the ranking model p. One may either seek to estimate the full ranking model p from degraded observations under a strong

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

structural assumption on p or else try to learn what is accessible from the observations and typically build an empirical ranking model, whose observable marginal distributions are close to those of the true underlying model.

Special attention has been paid in the literature to the situation where observations are *partial rankings* or *bucket orders*, defined as full rankings with ties, *i.e.* as orderings of the form $a_{1,1}, \ldots, a_{1,n_1} \prec \cdots \prec a_{r,1}, \ldots, a_{r,n_r}$ where $1 \leq r \leq n$ and $\sum_{i=1}^{r} n_i = n$ (see Critchlow, 1985; Fagin et al., 2006, for instance). This type of data includes most preferred items or more general top-k rankings. The information carried by such data is of "global" nature in the sense that it involves all the items (*e.g.* a most preferred item is preferred to every other item). Most estimation procedures originally designed for the observation of full rankings then easily extend to the case where available data are partial rankings (*e.g.* Lebanon & Lafferty, 2002; Lebanon & Mao, 2008; Huang et al., 2012; Kakarala, 2012).

In many practical situations however, the rank data at disposal only provide "local" information, in the sense that they do not involve all the n items but only small and variable subsets of items. In a recommendation setting for instance, users usually express their preferences through implicit feedback data related to the subset of recommended items, not to all the items in the catalog. The first step towards the analysis of such data is the analysis of orderings of the form $a_1 \succ \cdots \succ a_k$ with $k < n$, referred to as *incomplete rankings*. The case $k = 2$ corresponds to the setting of *pairwise comparisons*, the most widely considered in the literature (*e.g.* Wauthier et al., 2013; Busa-fekete et al., 2014; Rajkumar & Agarwal, 2014). Roughly speaking, the general problem of ranking model inference based on incomplete rankings is poorly understood. For instance, if the Placket-Luce model is naturally adapted to such rank data and can be used with various statistical estimation techniques (see Hunter, 2004; Guiver & Snelson, 2009; Azari Soufiani et al., 2013), inferring the Mallows model in this context is only possible with one method, introduced in (Lu & Boutilier, 2011). In addition, only two non-parametric methods capable of dealing with incomplete rankings of arbitrary length have been documented in the literature, those introduced in Kondor & Barbosa (2010) and Sun et al. (2012) namely.

It is the main purpose of this paper to propose a novel methodology for the statistical analysis of incomplete rankings. It crucially relies on a recent construction of a multiresolution analysis (MRA) for incomplete rankings, introduced in Clémençon et al. (2014). Related concepts and results involved in the subsequent analysis are summarized in section 3. The first contribution of the present article is the rigorous definition of an appropriate statistical framework in section 2, together with a full characterization of the components of a ranking model that can be statistically recovered without any structural assumption. Its second contribution is the introduction of a general method to learn these components in section 4. It is model-free and statistically efficient, while having tractable complexity, as shown in section 5. Experimental results are also provided in section 6 to illustrate its performance.

Notations. For a set E of finite cardinality $|E| < \infty$, we set $\mathcal{P}(E) = \{A \subset E \mid |A| \geq 2\}$ and denote by $L(E) = \{f : E \to \mathbb{R}\}$ the euclidean space of real-valued functions on E, equipped with the canonical inner product $\langle f, g \rangle = \sum_{x \in E} f(x)g(x)$ and the associated norm $\|f\| = (\sum_{x \in E} f(x)^2)^{1/2}$. The indicator function of a subset $S \subset E$ is denoted by $\mathbb{1}_S$ in general and by δ_x when S is the singleton $\{x\}$, in which case it is called a Dirac function. For a r.v. $\mathbf{X}$ and a probability distribution μ on a measurable space $\mathcal{X}$, the notation $\mathbf{X} \sim \mu$ means that μ is the probability distribution of $\mathbf{X}$. For any sigma-algebra $\mathcal{B}$, we denote by $\mathcal{F}(\mathcal{B}, \mathcal{X})$ the set of random variables on $\mathcal{X}$ that are $\mathcal{B}$-measurable.

2. Problem Statement

If full rankings can be regarded as permutations, it is more convenient to see an incomplete ranking $\pi_1 \succ \cdots \succ \pi_k$ as the *injective word* $\pi = \pi_1 \ldots \pi_k$, where $2 \leq k \leq n$ (see Kitaev, 2011). The content of a ranking π is the set $c(\pi) = \{\pi_1, \ldots, \pi_k\}$ and its length is the number $|\pi| = |c(\pi)|$. We denote by $\Gamma(A)$ the set of all injective words of content $A \in \mathcal{P}(\llbracket n \rrbracket)$ and by Γ_n the set of all incomplete rankings on $\llbracket n \rrbracket$. Notice that, equipped with these notations, $\Gamma(\llbracket n \rrbracket)$ corresponds to $\mathfrak{S}_n$. Word $\pi' \in \Gamma_n$ is a subword of word $\pi \in \Gamma_n$ if there exist indices $1 \leq i_1 < \cdots < i_{|\pi'|} \leq |\pi|$ such that $\pi' = \pi_{i_1} \ldots \pi_{i_{|\pi'|}}$, we then write $\pi' \subset \pi$.

Incomplete rankings - Probabilistic model. With these notations, the marginal distribution of a ranking model p on a subset $A \in \mathcal{P}(\llbracket n \rrbracket)$ is the probability distribution P_A over $\Gamma(A)$ defined for $\pi \in \Gamma(A)$ by

$$P_A(\pi) = \sum_{\sigma \in \mathfrak{S}_n,\, \pi \subset \sigma} p(\sigma). \tag{1}$$

As explained in the Introduction, it is assumed here that full realizations of a random permutation $\Sigma \sim p$ are not observable. The variability of rankings over a given subset $A \in \mathcal{P}(\llbracket n \rrbracket)$ is described by P_A, and it is thus natural to model the observed rankings over A as i.i.d. realizations of P_A. The complexity in learning from incomplete rankings stems from the fact that observations are not made on one single subset of items only, but on a possibly very large collection of (small) subsets. In e-commerce for instance, a user only expresses her preferences on the subset of items she came upon while browsing. We model the observation of an incomplete ranking by a random pair $(\mathbf{A}, \Pi)$,

where $\mathbf{A}$ is the subset of items involved in the ranking and Π is the ranking *per se*. We assume that $\mathbf{A}$ is independent from the underlying random variable Σ and drawn from an unknown probability distribution ν on $\mathcal{P}(\llbracket n \rrbracket)$. In this setting, a dataset $\mathcal{D}_N$ is formed of $N \geq 1$ i.i.d. pairs $(\mathbf{A}_1, \Pi^{(1)}), \ldots, (\mathbf{A}_N, \Pi^{(N)})$, drawn according to the following scheme:

$$\mathbf{A}_i \sim \nu \qquad \text{and} \qquad \Pi^{(i)} | (\mathbf{A}_i = A) \sim P_A. \tag{2}$$

Remark 1. The statistical model (2) can be viewed as a specific case of that introduced in Sun et al. (2012), where observations of incomplete rankings are modeled as censored observations of permutations, leading to the following setup: first a permutation Σ is drawn from p, then the subset of items $\mathbf{A}$ is drawn from a probability distribution ν_Σ on $\mathcal{P}(\llbracket n \rrbracket)$, and the ranking Π is taken equal to the ranking induced by Σ over $\mathbf{A}$. This setting boils down to ours when the distribution ν_σ is equal to ν for all $\sigma \in \mathfrak{S}_n$.

Identifiability. In this paper, the goal pursued is not to recover the true underlying model p. Focus is rather on learning an empirical ranking model as close as possible to the true underlying model p given ν. Indeed, all subsets of items are not assumed to be observable, which is the case in many applications where preferences are only observed on small subsets. In the absence of structural assumptions, the ranking model p is not identifiable. For instance, when only pairwise comparisons can be observed, one has access to the pairwise marginals, whose distributions are each characterized by a single parameter, leading to possibly $n(n-1)/2$ accessible parameters, whereas p is described by $n! - 1$ free parameters, and thus cannot be identified. In the general case, the accessible marginals are the P_A's for A lying in the support of the distribution ν, which is denoted by $\mathcal{A} = \{A \in \mathcal{P}(\llbracket n \rrbracket) \mid \nu(A) > 0\}$ and referred to as the *observation design* in the sequel. The number of accessible parameters is therefore equal to $\sum_{A \in \mathcal{A}} (|A|! - 1)$, which can be greater than $n! - 1$ even if $\llbracket n \rrbracket \notin \mathcal{A}$. It is then *a priori* unclear whether they charaterize p or, more generally, how many degrees of freedom they have. The answers are provided by the following theorem. For any collection of subsets $\mathcal{S} \subset \mathcal{P}(\llbracket n \rrbracket)$, we set $\mathcal{P}(\mathcal{S}) = \bigcup_{A \in \mathcal{S}} \mathcal{P}(A)$.

Theorem 1. *In absence of any restrictive assumption on the ranking model p, only the marginals p_B for $B \in \mathcal{P}(\mathcal{A})$ are identifiable from data drawn from (2). They are characterized by $\sum_{B \in \mathcal{P}(\mathcal{A})} d_{|B|}$ independent parameters, where d_k is the number of fixed-point free permutations (also called derangements) on a set with k elements.*

Theorem 1 is a direct consequence of the MRA of incomplete rankings introduced in Clémençon et al. (2014), briefly recalled in section 3. It shows for instance that if $\mathcal{A} = \{A \subset \llbracket n \rrbracket \mid 2 \leq |A| \leq k\}$ then the accessible parameters have $O(n^k)$ degrees of freedom. Attempting to estimate p without any assumption is thus vain in general.

Statistical framework. Practical applications require the construction of a ranking model, that can be used to compute probabilities of rankings on any subset of items $A \in \mathcal{P}(\llbracket n \rrbracket)$, as it is well illustrated in Sun et al. (2012). As $\Gamma_n = \bigsqcup_{A \in \mathcal{P}(\llbracket n \rrbracket)} \Gamma(A)$, we embed all the spaces $L(\Gamma(A))$ into $L(\Gamma_n)$. Let then $M_A : L(\Gamma_n) \to L(\Gamma(A))$ be the *marginal operator* on $A \in \mathcal{P}(\llbracket n \rrbracket)$ defined for any $f \in L(\Gamma_n)$ and $\pi \in \Gamma(A)$ by

$$M_A f(\pi) = \sum_{\sigma \in \Gamma_n,\ \pi \subset \sigma} f(\sigma),$$

so that $M_A p = P_A$ for all $A \in \mathcal{P}(\llbracket n \rrbracket)$ and $M_A f = 0$ if $f \in L(\Gamma(B))$ with $A \notin \mathcal{P}(B)$. We then consider the problem of building an empirical ranking model $\widehat{q}_N$ on $\mathfrak{S}_n$ from the dataset $\mathcal{D}_N$ such that its marginals $M_A \widehat{q}_N$ are accurate estimators of the P_A's, for $A \in \mathcal{P}(\mathcal{A})$.

Mathematically, building $\widehat{q}_N$ from $\mathcal{D}_N$ means that $\widehat{q}_N$ must be $\mathcal{B}_N$-measurable, where $\mathcal{B}_N$ is the σ-algebra generated by $\mathcal{D}_N$. We allow $\widehat{q}_N$ to take negative values but we impose that $\sum_{\sigma \in \mathfrak{S}_n} \widehat{q}_N(\sigma) = 1$. The possible negativity of $\widehat{q}_N(\sigma)$ for $\sigma \in \mathfrak{S}_n$ does not have much impact in practice because when the marginals $M_A \widehat{q}_N$ are close to the P_A's, they only take positive values. We evaluate the quality of the estimation of P_A for $A \in \mathcal{A}$ through the mean squared error (MSE) $\mathbb{E}[\|M_A \widehat{q}_N - P_A\|_A^2]$, where $\|\cdot\|_A$ denotes the euclidean norm on $L(\Gamma(A))$. Here and throughout the article, the symbol $\mathbb{E}$ represents the expectation with respect to the randomly drawn dataset $\mathcal{D}_N$. As we consider the problem of simultaneous estimation of the marginals, it is natural to consider the sum of the errors on each A weighted by ν.

Definition 1 (Performance measure)**.** The performance of an empirical ranking model $\widehat{q}_N \in \mathcal{F}(\mathcal{B}_N, L(\mathfrak{S}_n))$ with $\sum_{\sigma \in \mathfrak{S}_n} \widehat{q}_N(\sigma) = 1$ is given by

$$\mathcal{E}(\widehat{q}_N) := \sum_{A \in \mathcal{A}} \nu(A) \mathbb{E}\left[\|M_A \widehat{q}_N - P_A\|_A^2\right].$$

Statistical and computational challenge. We emphasize that constructing an accurate empirical ranking model $\widehat{q}_N$ in this setting is far from being trivial, because the marginals $M_A \widehat{q}_N$ are linked through highly entangled combinatorial relationships. Estimating marginals on different subsets thus cannot be done separately (refer to the Supplementary Material for an illustrative example). Even if the P_A's are assumed to be known, finding a function $q \in L(\mathfrak{S}_n)$ such that $M_A q = P_A$ for all $A \in \mathcal{A}$ requires to solve a linear system of $\sum_{A \in \mathcal{A}} |A|!$ equations with $n!$ unknowns, where each equation $M_A q(\pi) = P_A(\pi)$ for $A \in \mathcal{A}$ and $\pi \in \Gamma(A)$ involves at least $n!/|A|!$ operations. It therefore represents a daunting computational task as soon as $n > 10$, whereas n is around 10^4 in practical applications. Fortunately, as shall be seen below, the MRA framework brings a new representation of the data tailored to this task, drastically reducing this complexity.

3. MRA of Incomplete Rankings

Multiresolution analysis of incomplete rankings crucially relies on the following result: any function f on $\mathfrak{S}_n$ can be decomposed as a sum of components that each localize the specific information of one marginal $M_B f$ for $B \in \mathcal{P}(\llbracket n \rrbracket)$, in the sense that the marginal $M_A f$ on any subset $A \in \mathcal{P}(\llbracket n \rrbracket)$ only involves the components specific to the subsets $B \in \mathcal{P}(A)$. This is formalized in the following theorem, established in Clémençon et al. (2014). Let us denote by $V^0 = \mathbb{R}\mathbb{1}_{\mathfrak{S}_n}$ the 1-d space of constant functions in $L(\mathfrak{S}_n)$, and for $f \in L(\Gamma_n)$, we set $\bar{f} = \sum_{\pi \in \Gamma_n} f(\pi)$.

Theorem 2. *There exists a collection $(W_B)_{B \in \mathcal{P}(\llbracket n \rrbracket)}$ of subspaces of $L(\mathfrak{S}_n)$ orthogonal to V^0 such that:*

1. $L(\mathfrak{S}_n) = V^0 \oplus \bigoplus_{B \in \mathcal{P}(\llbracket n \rrbracket)} W_B$,
2. *For $B \in \mathcal{P}(\llbracket n \rrbracket)$ and $f \in W_B$, $M_B f = 0 \Rightarrow f = 0$,*
3. *For $A, B \in \mathcal{P}(\llbracket n \rrbracket)$ with $B \not\subset A$ and $f \in W_B$, $M_A f = 0$.*

In particular for $f \in L(\mathfrak{S}_n)$, decomposed as $f = (1/n!)\bar{f} + \sum_{B \in \mathcal{P}(\llbracket n \rrbracket)} \tilde{f}_B$, and $A \in \mathcal{P}(\llbracket n \rrbracket)$, one has:

$$M_A f = \frac{1}{|A|!}\bar{f} + \sum_{B \in \mathcal{P}(A)} M_A \tilde{f}_B.$$

At last, $\dim W_B = d_{|B|}$ for all $B \in \mathcal{P}(\llbracket n \rrbracket)$.

Multiscale structure. MRA allows to exploit the natural multiscale structure of the marginals $M_A f$ of any function $f \in L(\mathfrak{S}_n)$. Here, the notion of *scale* corresponds to the number of items in the subset on which the marginal is considered. For $k \in \{2, \ldots, n\}$, the space of scale k is defined by $W^k = \bigoplus_{|B|=k} W_B$. One thus obtains $L(\mathfrak{S}_n) = V^0 \oplus \bigoplus_{k=2}^n W^k$. This last decomposition of $L(\mathfrak{S}_n)$ is somehow analogous to classic MRA on $L^2(\mathbb{R})$ and offers a similar interpretation: if a function $f \in L(\mathfrak{S}_n)$ is projected onto $V^0 \oplus \bigoplus_{k=2}^K W^k$ with $K \in \{2, \ldots, n\}$, then only the information of f of scale up to K can be captured. In other words, starting from a constant function in V^0, each space W^k provides the supplementary level of details specific to scale k. The decomposition of $L(\mathfrak{S}_n)$ given by Theorem 2 actually allows to further decompose a function $f \in L(\mathfrak{S}_n)$ in the "space of items", where each component $\tilde{f}_B$ provides the supplementary level of details specific to the marginal $M_B f$, for $B \in \mathcal{P}(\llbracket n \rrbracket)$. To a certain extent, this "space-scale" localization is analogous to the classic space-scale localization in wavelet analysis.

When incomplete ranking data generated through the scheme (2) are observed, one may form empirical versions of the marginals $M_A p$ of the ranking model p on the subsets $A \in \mathcal{A}$ and represent them as elements of $L(\Gamma(A))$. This raw representation of the data is however not efficient for a statistical analysis because it does not allow to exploit the structure induced by equation (1). In contrast, the MRA framework brings a new representation that defines efficient "features" for statistical analysis.

To define this representation, we enter the details of the construction MRA framework. For $B \in \mathcal{P}(\llbracket n \rrbracket)$, let H_B be the subspace of $L(\Gamma(B))$ defined by

$$H_B = \{F \in L(\Gamma(B)) \mid M_{B'} F = 0 \text{ for all } B' \subsetneq B\}.$$

Let $\bar{0}$ be by convention the unique injective word of content $\emptyset$ and length 0. Word $\pi' \in \Gamma_n$ is said a *contiguous subword* of word $\pi \in \Gamma_n$ if there exists $i \in \{1, \ldots, |\pi| - |\pi'| + 1\}$ such that $\pi' = \pi_i \pi_{i+1} \ldots \pi_{i+|\pi'|-1}$. This is denoted by $\pi' \sqsubset \pi$. For $A \in \mathcal{P}(\llbracket n \rrbracket)$, define the linear embedding operator $\phi_A : L(\Gamma_n \cup \{\bar{0}\}) \to L(\Gamma(A))$ for $\sigma \in \Gamma(A)$ by $\phi_A \delta_{\bar{0}}(\sigma) = 1/|A|!$ and for any ranking $\pi \in \Gamma_n$ by

$$\phi_A \delta_\pi(\sigma) = \frac{1}{(|A| - |\pi| + 1)!} \quad \text{if } \pi \sqsubset \sigma \text{ and 0 otherwise.}$$

Then for $A \in \mathcal{P}(\llbracket n \rrbracket)$, $\phi_{\llbracket n \rrbracket}$ is an isomorphism between H_A and W_A. This means that for $f \in L(\mathfrak{S}_n)$ and $B \in \mathcal{P}(\llbracket n \rrbracket)$, there exists a unique element $X_B^{\llbracket n \rrbracket} f \in H_B$ such that $\tilde{f}_B = X_B^{\llbracket n \rrbracket} f$. Setting $X_\emptyset f = \bar{f} \delta_{\bar{0}}$, one has

$$f = \sum_{B \in \mathcal{P}(\llbracket n \rrbracket) \cup \{\emptyset\}} \phi_{\llbracket n \rrbracket} X_B^{\llbracket n \rrbracket} f.$$

More generally, Theorem 2 still holds true for any space $L(\Gamma(A))$ with $A \in \mathcal{P}(\llbracket n \rrbracket)$, so that $L(\Gamma(A)) = \mathbb{R}\mathbb{1}_{\Gamma(A)} \oplus \bigoplus_{B \in \mathcal{P}(A)} W_B^A$, where for $B \in \mathcal{P}(A)$, W_B^A is the detail subspace of $L(\Gamma(A))$ related to B, and ϕ_A is an isomorphism between H_B and W_B^A. Thus for any function $F \in L(\Gamma(A))$, there exists a unique family $(X_B^A F)_{B \in \mathcal{P}(A)}$ with $X_B^A F \in H_B$ for each $B \in \mathcal{P}(A)$ such that

$$F = \sum_{B \in \mathcal{P}(A) \cup \{\emptyset\}} \phi_A X_B^A F,$$

where $X_\emptyset^A F = \bar{F} \delta_{\bar{0}}$. This defines for any $B \in \mathcal{P}(\llbracket n \rrbracket)$ the operator $X_B : L(\Gamma_n) \to H_B$ on each space $L(\Gamma(A))$ with $A \in \mathcal{P}(\llbracket n \rrbracket)$ by

$$X_B F = X_B^A F \text{ if } B \subset A \text{ and 0 otherwise,}$$

for $F \in L(\Gamma(A))$. Now, a result from Clémençon et al. (2014) shows that for any $A, A', B \in \mathcal{P}(\llbracket n \rrbracket)$ such that $B \subset A' \subset A$ and any $F \in L(\Gamma(A))$,

$$X_B M_{A'} F = X_B F. \tag{3}$$

Equation (3) means that the component $X_B F$ of F is "invariant under marginal transform". In a statistical learning perspective, the marginals should thus be more efficiently estimated when represented through the projectors X_B. We call the operators X_B the *wavelet projectors*, and define the *wavelet transform* as the operator

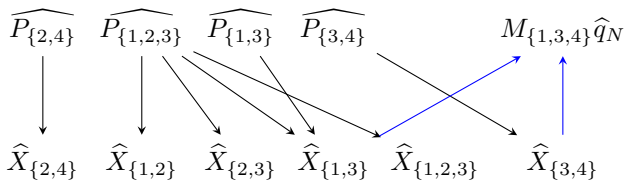

Figure 1. Principle of the MRA-based linear ranking model

$\Psi_X : L(\Gamma_n) \to \bigoplus_{B\in\mathcal{P}(\llbracket n\rrbracket)} H_B, F \mapsto (X_B F)_{B\in\mathcal{P}(\llbracket n\rrbracket)}$. The wavelet transform defines the representation that we use for any function $F \in L(\Gamma(A))$ with $A \in \mathcal{P}(\llbracket n\rrbracket)$. We summarize its properties in the following theorem.

Theorem 3 (MRA representation)**.** *Let $A \in \mathcal{P}(\llbracket n\rrbracket)$ and $F \in L(\Gamma(A))$. Then*

$$F = \sum_{B\in\mathcal{P}(A)\cup\{\emptyset\}} \phi_A X_B F \quad \textit{and}$$

$$M_{A'}F = \sum_{B\in\mathcal{P}(A')\cup\{\emptyset\}} \phi_{A'} X_B F \quad \textit{for any } A' \in \mathcal{P}(A).$$

4. MRA-based Estimation

The MRA framework shows that the marginals P_A for $A \in \mathcal{A}$ of the ranking model p are characterized by the wavelet projections $X_B p$ for $B \in \mathcal{P}(\mathcal{A}) := \bigcup_{A\in\mathcal{A}} \mathcal{P}(A)$. This proves Theorem 1 and shows more specifically that the component of p that can be learned when data is drawn from the scheme (2) is $\phi_{\llbracket n\rrbracket} \sum_{B\in\mathcal{P}(\mathcal{A})} X_B p$. Based on this observation, we introduce a general learning method that relies on the estimation of the wavelet projections $X_B p$ of the ranking model p for all $B \in \mathcal{P}(\mathcal{A})$.

Definition 2 (MRA-based empirical ranking model)**.** The general MRA-based learning method consists in constructing for each $B \in \mathcal{P}(\mathcal{A})$ an estimator $\widehat{X}_B \in \mathcal{F}(\mathcal{B}_N, H_B)$ of $X_B p$ from the dataset $\mathcal{D}_N$. The empirical ranking model is then given by $\widehat{q}_N = \sum_{B\in\mathcal{P}(\mathcal{A})\cup\{\emptyset\}} \phi_{\llbracket n\rrbracket} \widehat{X}_B$, where we fix $\widehat{X}_\emptyset = \delta_{\bar{0}}$ for all empirical ranking models.

From dependence to independence. We underline that the major advantage of this approach consists in the fact that, as the wavelet projections are by construction linearly independent, it turns the complex problem of simultaneously estimating the marginals into a collection of estimation problems that can be solved independently. The performance of a ranking model constructed this way can be easily controlled via the following proposition.

Proposition 1. *Let $(\widehat{X}_B)_{B\in\mathcal{P}(\mathcal{A})}$ with $\widehat{X}_B \in \mathcal{F}(\mathcal{B}_N, H_B)$ for each $B \in \mathcal{P}(\mathcal{A})$ and $\widehat{q}_N$ the associated empirical ranking model from definition 2. Then*

$$\mathcal{E}(\widehat{q}_N) \leq \sum_{B\in\mathcal{P}(\mathcal{A})} \nu_\phi(B)\mathbb{E}\left[\left\|\widehat{X}_B - X_B p\right\|_B^2\right]$$

where $\nu_\phi(B) = \sum_{A\subset\llbracket n\rrbracket,\ B\subset A} \nu(A) 2^{|A|}/(|A| - |B| + 1)!$.

Sketch of proof. One can show that for $A, B \in \mathcal{P}(\llbracket n\rrbracket)$ with $B \subset A$ and $F \in L(\Gamma(B))$, $\|\phi_A F\|_A^2 = \|F\|_B^2/(|A| - |B| + 1)!$. Then the proof is concluded using Theorem 3 and the Cauchy-Schwarz inequality. Refer to the Supplementary Material for the technical details. □

Among all the empirical ranking models that can be built according to the principle formulated in Definition 2, we consider a specific class of models, that shall be referred to as MRA-based linear empirical ranking models. Its definition is based on several empirical estimates built from the dataset $\mathcal{D}_N = \{(\mathbf{A}_1, \Pi^{(1)}), \ldots, (\mathbf{A}_N, \Pi^{(N)})\}$. For $A \in \mathcal{P}(\llbracket n\rrbracket)$, we set $\widehat{I}_A = \{1 \leq i \leq N \mid \mathbf{A}_i = A\}$, so that "$\widehat{I}_A \neq \emptyset$" means that A is observed in the dataset $\mathcal{D}_N$. The empirical estimator $\widehat{\nu}_N$ of ν is naturally defined by $\widehat{\nu}_N(A) = |\widehat{I}_A|/N$ for $A \in \mathcal{P}(\llbracket n\rrbracket)$. We denote its support by $\widehat{\mathcal{A}}_N$ and refer to it as the *empirical observation design*. Notice that $\widehat{\mathcal{A}}_N \subset \mathcal{A}$, the support of ν. For $A \in \mathcal{P}(\llbracket n\rrbracket)$, we define the empirical estimator of P_A for $\pi \in \Gamma(A)$ by $\widehat{P_A}(\pi) = |\{i \in \widehat{I}_A \mid \Pi^{(i)} = \pi\}/|\widehat{I}_A|$ if $A \in \widehat{\mathcal{A}}_N$ and $1/|A|!$ otherwise. We denote by $\mathcal{B}_N^\nu$ the σ-algebra generated by $\widehat{\nu}_N$.

For a given subset of items $B \in \mathcal{P}(\mathcal{A})$, there are two possibilities. Either $B \notin \mathcal{P}(\widehat{\mathcal{A}}_N)$, meaning that it is not included in any of the observed sets of items. In this case one cannot infer anything about $\widehat{X}_B$ without additional regularity assumption. Or else $B \in \mathcal{P}(\widehat{\mathcal{A}}_N)$, meaning that there exists at least one observed subset of items $A \in \widehat{\mathcal{A}}_N$ such that $B \subset A$. Then natural candidates for $\widehat{X}_B$ are the wavelet projections $X_B\widehat{P_A}$ of the empirical estimators $\widehat{P_A}$ for $A \in \widehat{\mathcal{A}}_N$ such that $B \subset A$. This essential observation motivates the learning method proposed below. For $B \in \mathcal{P}(\llbracket n\rrbracket)$, we set $\mathcal{Q}(B) = \{A \subset \llbracket n\rrbracket \mid B \subset A\}$.

Definition 3 (MRA-based linear ranking model)**.** For $B \in \mathcal{P}(\mathcal{A})$ and $\widehat{\theta} \in \mathcal{F}(\mathcal{B}_N, \mathbb{R}^{2^n})$, the MRA-based linear estimator related to the weighting vector $\widehat{\theta}$ is defined by

$$\widehat{X}_{B,\widehat{\theta}} = \sum_{A\in\widehat{\mathcal{A}}_N\cap\mathcal{Q}(B)} \widehat{\theta}(A) X_B \widehat{P_A},$$

where by convention $\widehat{X}_{B,\widehat{\theta}} = 0$ when $\widehat{\mathcal{A}}_N \cap \mathcal{Q}(B) = \emptyset$ or equivalently $B \notin \mathcal{P}(\widehat{\mathcal{A}}_N)$. We denote by $\widehat{q}_{N,\widehat{\theta}}$ the related empirical ranking model and fix $\widehat{X}_{\emptyset,\widehat{\theta}} = \delta_{\bar{0}}$.

Example 1. The principle of the MRA-based estimator is depicted in Fig. 1. In this example, we assume that $\widehat{\mathcal{A}}_N = \{\{2,4\},\{1,2,3\},\{1,3\},\{3,4\}\}$. For each $A \in \widehat{\mathcal{A}}_N$, the empirical estimator $\widehat{P_A}$ contributes to the estimators $\widehat{X}_B$ of $X_B p$ for all $B \in \mathcal{P}(A)$. Then the prediction on a (possibly unobserved) subset A, $A = \{1,3,4\}$ in the illustration, only involves the $\widehat{X}_B$'s for $B \in \mathcal{P}(A) \cap \mathcal{P}(\widehat{\mathcal{A}}_N)$.

Remark 2. Several methods for ranking aggregation or estimation rely on the breaking of rankings into pairwise comparisons (see Hüllermeier et al., 2008, for instance). For the usual parametric ranking models, it is usually shown that these methods do not degrade the available information too much (see for instance Lu & Boutilier, 2011; Meek & Meila, 2014). When no structural assumption is made however on the ranking model p, breaking all the observed rankings into pairwise comparisons boils down to suppressing all the information of scale higher than 2 defined in the MRA framework. In particular any MRA-based linear ranking model defined from pairwise comparisons only would be such that $\widehat{X}_{B,\widehat{\theta}} = 0$ for all $B \subset [\![n]\!]$ with $|B| > 2$.

The following result provides asymptotic guarantees for the accuracy of the MRA-based empirical ranking model we proposed. We say that an empirical ranking model $\widehat{q}_N \in \mathcal{F}(\mathcal{B}_N, L(\mathfrak{S}_n))$ is asymptotically unbiased if $\lim_{N\to\infty} \mathbb{E}[M_A \widehat{q}_N] = P_A$ for all $A \in \mathcal{A}$.

Proposition 2. *Let $\widehat{\theta} \in \mathcal{F}(\mathcal{B}_N^\nu, \mathbb{R}^{2^n})$. Then the MRA-based linear ranking model $\widehat{q}_{N,\widehat{\theta}}$ is asymptotically unbiased if $\lim_{N\to\infty} \mathbb{E}\left[\sum_{A\in\widehat{\mathcal{A}}_N\cap\mathcal{Q}(B)} \widehat{\theta}(A)\right] = 1$ for all $B \in \mathcal{P}(\mathcal{A})$.*

Sketch of proof. By Theorem 3, one has $\mathbb{E}[M_A\widehat{q}_N] = \sum_{B\in\mathcal{P}(A)\cup\{\emptyset\}} \phi_A \mathbb{E}[\widehat{X}_{B,\widehat{\theta}}]$ for $A \in \mathcal{A}$. Now, for $B \in \mathcal{P}(A)$ and $\pi \in \Gamma(B)$, one can show after some calculations that $\mathbb{E}[\widehat{X}_{B,\widehat{\theta}}(\pi)] = X_B p(\pi) \mathbb{E}\left[\sum_{A\in\widehat{\mathcal{A}}_N\cap\mathcal{Q}(B)} \widehat{\theta}(A)\right]$. This suffices to conclude the proof. Refer to the Supplementary Material for the technical details. □

Notice that Proposition 2 requires that the weights $\widehat{\theta}(A)$ are $\mathcal{B}_N^\nu$-measurable, in other words that they are constructed from $\widehat{\nu}_N$. They can be constructed from $\widehat{\mathcal{A}}_N$ but not from the $\widehat{P_A}$'s for $A \in \widehat{\mathcal{A}}_N$ for instance. This hypothesis is however not too limiting in practice. It is satisfied in particular by the *weighted least square estimator* defined below.

Definition 4 (WLS estimator)**.** Let $B \in \mathcal{P}(\mathcal{A})$. Given $\widehat{\nu}_N$, the solutions of the following minimization problem

$$\min_{\widehat{\theta}\in\mathbb{R}^{2^n}} \sum_{A\in\widehat{\mathcal{A}}_N\cap\mathcal{Q}(B)} \widehat{\nu}_N(A) \|\widehat{X}_{B,\widehat{\theta}} - X_B \widehat{P_A}\|_B^2$$

are the vectors $\widehat{\theta} \in \mathbb{R}^{2^n}$ defined for all $A \in \widehat{\mathcal{A}}_N \cap \mathcal{Q}_B$ by

$$\widehat{\theta}^{WLS}(A) := \frac{\widehat{\nu}_N(A)}{\sum_{A'\in\widehat{\mathcal{A}}_N\cap\mathcal{Q}(B)} \widehat{\nu}_N(A')}. \tag{4}$$

We then define the weighted least square estimator by $\widehat{X}_B^{WLS} := \widehat{X}_{B,\widehat{\theta}^{WLS}}$ and denote the related empirical ranking model by $\widehat{q}_N^{WLS}$.

Beyond the fact that the WLS ranking model is a natural choice among the class of MRA-based linear empirical ranking models, the result stated below reveals that it is asymptotically unbiased with an error rate of order $O(1/N)$, the optimal rate in parametric estimation.

Theorem 4. *The WLS estimator $\widehat{q}_N^{WLS}$ is asymptotically unbiased and has an error bounded by*

$$\mathcal{E}\left(\widehat{q}_N^{WLS}\right) \leq \frac{C_1}{N} + C_2\rho^{2N}$$

for all $N \geq 1$, where $0 < \rho < 1$ is a constant that only depends on ν and C_1 and C_2 are positive constants that only depend on p and ν, given in the Supplementary Material.

Sketch of proof. First, using (4) one obtains, for $B \in \mathcal{P}(\mathcal{A})$, $\mathbb{E}\left[\sum_{A\in\widehat{\mathcal{A}}_N\cap\mathcal{Q}(B)} \widehat{\theta}^{WLS}(A)\right] = 1 - (1 - \sum_{A\in\mathcal{Q}(B)} \nu(A))^N$. Since $B \in \mathcal{P}(A)$, $\mathcal{A} \cap \mathcal{Q}(B) \neq \emptyset$ and $\sum_{A\in\mathcal{Q}(B)} \nu(A) > 0$. Then Proposition 2 ensures that $\widehat{q}_N^{WLS}$ is asymptotically unbiased. To bound the error, we use Proposition 1 and calculate explicitly the terms $\mathbb{E}\left[\|\widehat{X}_B^{WLS} - X_B p\|_B^2\right]$ for $B \in \mathcal{P}(\mathcal{A})$ through the bias-variance decomposition. Refer to the Supplementary Material for technical details. □

Remark 3. The two non-parametric approaches proposed in the literature to handle incomplete rankings in Kondor & Barbosa (2010) and Sun et al. (2012) both rely on kernel regularization of a global estimator defined as

$$\widehat{p}_N = \frac{1}{N}\sum_{i=1}^N \frac{|\mathbf{A}_i|!}{n!} \mathbb{1}_{\mathfrak{S}_n(\Pi^{(i)})} \tag{5}$$

in the present setting, where $\mathfrak{S}_n(\pi) = \{\sigma \in \mathfrak{S}_n \mid \pi \subset \sigma\}$ is the set of linear extensions of the ranking $\pi \in \Gamma_n$. The choice of this estimator relies on the following heuristic: the observation of an incomplete ranking π is actually a degraded observation of a full ranking $\sigma \in \mathfrak{S}_n(\pi)$. Thus it gives the same information as the uniform distribution over $\mathfrak{S}_n(\pi)$, equal to $(|\pi|!/n!)\mathbb{1}_{\mathfrak{S}_n(\pi)}$. The estimator $\widehat{p}_N$ is then simply the average of the uniform distributions over the $\mathfrak{S}_n(\Pi^{(i)})$'s. Though this assumption is appealing, the estimator $\widehat{p}_N$ is fundamentally biased. Indeed for $B \in \mathcal{P}([\![n]\!])$ and $\pi \in \Gamma(B)$, $\mathbb{E}\left[M_B \widehat{p}_N(\pi)\right] =$

$$\sum_{A\in\mathcal{A}} \nu(A)\frac{|A|!}{n!} \sum_{\pi'\in\Gamma(A)} P_A(\pi') \left\langle \mathbb{1}_{\mathfrak{S}_n(\pi')}, \mathbb{1}_{\mathfrak{S}_n(\pi)} \right\rangle,$$

which is usually different from $P_B(\pi)$ because the $\mathfrak{S}_n(\pi)$'s are not disjoint for different π's in general. For instance in the case where only pairwise comparisons are observed, one obtains after some calculations, for $i \neq j \in [\![n]\!]$:

$$\mathbb{E}\left[M_{\{i,j\}}\widehat{p}_N(ij)\right] = \frac{1}{2} + \nu(\{i,j\})P'_{\{i,j\}}(ij)$$
$$+\frac{1}{3}\sum_{k\in[\![n]\!]\setminus\{i,j\}} \left[\nu(\{i,k\})P'_{\{i,k\}}(ik) + \nu(\{j,k\})P'_{\{j,k\}}(kj)\right],$$

where for $k \neq l \in [\![n]\!]$ and $\pi \in \Gamma(\{k,l\})$, $P'_{\{k,l\}}(kl) := P_{\{k,l\}}(kl) - 1/2$. Except if the distribution ν is concentrated on the pair $\{i,j\}$ solely (in which case there is not much to say), this expression shows that the estimate $\mathbb{E}\left[M_{\{i,j\}}\widehat{p}_N(ij)\right]$ blends the probabilities of many other pairwise comparisons and is therefore fundamentally different from $P_{\{i,j\}}(ij)$.

The MRA representation allows to exploit only the information from the observed dataset, which is why the constants C_1 and C_2 in Theorem 4 depend on p and ν and not directly on n. We point out however that the more diffuse ν is, the more degrees of freedom the dataset has, and the bigger they are. The same interpretation applies to the computational aspects.

5. Computational Advantages

To be useful in practice, an empirical ranking model $\widehat{q}_N \in L(\mathfrak{S}_n)$ must face the following three intertwined computational challenges.

1. **Storage of the ranking model:** the naive storage of a vector $(\widehat{q}_N(\sigma))_{\sigma\in\mathfrak{S}_n}$ requires $n! - 1$ parameters, which is largely unfeasible when n becomes greater than 15.
2. **Complexity of the learning procedure:** the learning procedure can require a drastic amount of operations, because of the high dimensionality of the data.
3. **Complexity of the computation of a marginal:** for $A \in \mathcal{P}([\![n]\!])$ and $\pi \in \Gamma(A)$, the naive computation of $M_A\widehat{q}_N(\pi)$ for $\pi \in \Gamma(A)$ requires $n!/|A|!$ operations.

Example 2. Consider the empirical model $\widehat{p}_N$ defined in Remark 3 by (5). It can be rewritten as

$$\widehat{p}_N = \sum_{A\in\widehat{\mathcal{A}}_N} \widehat{\nu}_N(A) \sum_{\pi\in\Gamma(A)} \widehat{P_A}(\pi)\mathbb{1}_{\mathfrak{S}_n(\pi)}.$$

Its most efficient storage is under the form of the collections of parameters $(\widehat{\nu}_N(A))_{A\in\widehat{\mathcal{A}}_N}$ and $(\widehat{P_A}(\pi))_{A\in\widehat{\mathcal{A}}_N,\,\pi\in\Gamma(A)}$, and the learning procedure is naturally in $O(N)$. But then, each computation of the marginal probability of a ranking $\pi' \in \Gamma_n$ involves the computation of all the inner products $\left\langle \mathbb{1}_{\mathfrak{S}_n(\pi')}, \mathbb{1}_{\mathfrak{S}_n(\pi)} \right\rangle$ for $\pi \in \bigsqcup_{A\in\widehat{\mathcal{A}}_N} \Gamma(A)$. This is at the root of the main computational limitation of the approaches introduced in Kondor & Barbosa (2010) and Sun et al. (2012).

The MRA-based linear empirical ranking model is defined in terms of wavelet projections and thus naturally overcomes challenges 1. and 3., as shown by the following result (see the Supplementary Material for the technical proof). We denote by $K = \max_{A\in\mathcal{A}} |A|$ the maximum size of an observable subset.

Proposition 3. *Let $\widehat{q}_N \in \mathcal{F}(\mathcal{B}_N, L(\mathfrak{S}_n))$ be a MRA-based linear ranking model.*

- *Its storing requires a number of parameters upper bounded by $K!\, 2^K \min(N, |\mathcal{A}|)$.*
- *The computation of $M_A\widehat{q}_N(\pi)$ for $A \in \mathcal{P}([\![n]\!])$ and $\pi \in \Gamma(A)$ needs less than $|A|(|A|-1)/2$ operations.*

As K is small in practical applications, these bounds are quite reasonable. Analyzing the complexity of the learning procedure for a general MRA-based linear ranking model is too intricate because it depends on the choice of weighting vector $\widehat{\theta} \in \mathbb{R}^{2^n}$. We do it for the weighted least squares model, using the following explicit formula (proved in the Supplementary Material).

Proposition 4. *For $B \in \mathcal{P}(\widehat{\mathcal{A}}_N)$ and $\pi \in \Gamma(B)$,*

$$\tilde{X}_B^{WLS}(\pi) = \frac{\sum_{\pi'\in\Gamma(B)} \alpha_B(\pi,\pi')|\{1 \leq i \leq N \mid \pi' \subset \Pi^{(i)}\}|}{\sum_{A\in\widehat{\mathcal{A}}_N\cap\mathcal{Q}(B)} |I_A|},$$

where $\alpha_B(\pi,\pi') := X_B\delta_\pi(\pi')$ for $\pi' \in \Gamma(B)$.

The coefficients $\alpha_B(\pi,\pi')$ for $B \in \mathcal{P}([\![n]\!])$ and $\pi,\pi' \in \Gamma(B)$ do not depend on the application nor on the dataset, they can be pre-computed. It is shown in the Supplementary Material how their computation can be implemented with complexity bounded by $K^2K!$.

Proposition 5. *Assuming that all coefficients $\alpha_B(\pi,\pi')$ for $B \in \mathcal{P}(\mathcal{A})$ and $\pi,\pi' \in \Gamma(B)$ have been pre-computed, the WLS ranking model can be learned with complexity bounded by $2^K(K!+1)(N+|\mathcal{A}|)$.*

Refer to the Supplementary Material for the proof of Proposition 5. The bounds given by Propositions 3 and 5 are not small but they depend directly on the complexity of the observed dataset, not on the number of items n. This is a great achievement regarding the computational challenges of the analysis of incomplete rankings.

6. Numerical Experiments

Here we examine the performance of the WLS empirical ranking model in numerical experiments and compare it with three others: the Plackett-Luce model (estimated by means of the MM algorithm introduced in Hunter (2004)), the estimator from Sun et al. (2012), called SLK (we take the bandwidth of the kernel equal to $\binom{n}{2} + 1$ to be sure that the smoothing is applied to the entire dataset), and the collection of empirical estimators $(\widehat{P_A})_{A\in\mathcal{A}}$. The latter are not given as the marginals of a ranking model and besides, it may be the case that no ranking model induces them due to sampling noise. It is however interesting to see them as a baseline.

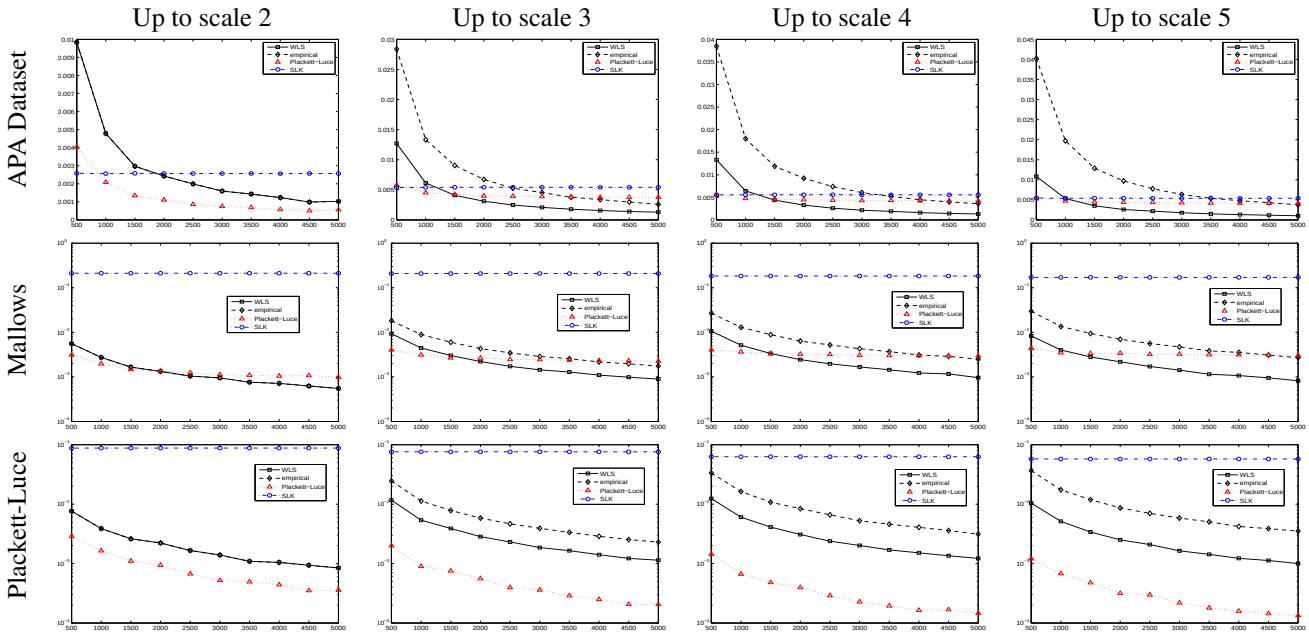

Figure 2. Evolution of the performance $\mathcal{E}(\widehat{q}_N)$ with N for each estimator: WLS in squares, empirical in diamonds, Plackett-Luce in triangles and SLK in circles, with different underlying ranking models: APA dataset (first row), Mallows (second row), Plackett-Luce (third row) and with probability ν uniform on $\{A \subset [\![5]\!] \mid 2 \leq |A| \leq k\}$ with $k = 2, 3, 4, 5$ (from left to right). For the Mallows and Plackett-Luce models, the results are represented on a logarithmic scale.

Each experiment is characterized by a ranking model p, a probability distribution ν and a number of observations N. We consider two theoretical ranking models, namely a Plackett-Luce model defined for $\sigma \in \mathfrak{S}_n$ by $p_{\mathbf{w}}(\sigma) = \prod_{i=1}^n w_{\sigma_i} / (\sum_{j=i}^n w_{\sigma_j})$ with parameter vector $\mathbf{w} = (w_1, \ldots, w_n)$ drawn uniformly at random on the simplex $\{\mathbf{x} \in [0,1]^n \mid \sum_{i=1}^n x_i = 1\}$ and a Mallows model defined for $\sigma \in \mathfrak{S}_n$ by $p(\sigma) \propto e^{-T(\sigma_0, \sigma)}$ where T is the Kendall's tau distance on $\mathfrak{S}_n$ and $\sigma_0 = 12 \ldots n$, and one empirical model, namely the distribution of the 5738 votes in the APA dataset (see Diaconis, 1989) that we consider as a ground truth ranking model. In all the experiments, $n = 5$. For each ranking model, we examine the four different settings where ν is the uniform probability distribution on $\{A \subset [\![5]\!] \mid 2 \leq |A| \leq k\}$ for $k = 2, 3, 4, 5$, and let the size of the drawn dataset $\mathcal{D}_N$ vary between 500 and 5000. We then evaluate the performance of an empirical ranking model $\widehat{q}_N$ constructed from $\mathcal{D}_N$ through a Monte-Carlo estimate of the performance $\mathcal{E}(\widehat{q}_N)$ averaged from 100 drawings of $\mathcal{D}_N$.

Fig. 2 depicts the experimental results. As explained in Remark 3, the SLK ranking model applies a strong smoothing which leads to a very small variance but an important bias when p differs from the uniform distribution on $\mathfrak{S}_n$. This is why it converges rapidly and its performance is almost constant through the experiments for $N \geq 500$. The Plackett-Luce model relies on a structural assumption and is thus naturally biased when p is not a Plackett-Luce model. This explains why it does not perform best in the latter case. As noticed in Remark 2, the marginals of the WLS ranking model are equal to the empirical estimators when only pairwise comparisons are observed. More generally, they are both asymptotically unbiased whatever the underlying ranking model p and have similar behaviors, except that the WLS has reduced variance and thus converges faster. Globally, the WLS ranking model quickly outperforms its competitors when N grows.

7. Conclusion

In this paper, we rigorously formulated the issue of learning a ranking model from incomplete rankings, as the problem of building an empirical ranking model whose marginals are good estimators of those of the true underlying ranking model, when they are observable. Based on the concept of MRA of incomplete rankings introduced in (Clémençon et al., 2014), we provided a general framework to construct an asymptotically unbiased ranking model with an optimal convergence rate, which can be computed with complexity directly depending on the data. These theoretical guarantees, as well as the good performance observed in numerical experiments, are an encouragement to pursue the development of this framework, for instance to define efficient regularization procedures or apply it to other statistical tasks.

References

Azari Soufiani, Hossein, Chen, William, Parkes, David C, and Xia, Lirong. Generalized method-of-moments for rank aggregation. In *Advances in Neural Information Processing Systems 26*, pp. 2706–2714, 2013.

Busa-fekete, Robert, Huellermeier, Eyke, and Szrnyi, Balzs. Preference-based rank elicitation using statistical models: The case of mallows. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1071–1079, 2014.

Clémençon, Stéphan, Jakubowicz, Jérémie, and Sibony, Eric. Multiresolution analysis of incomplete rankings. *ArXiv e-prints*, march 2014.

Critchlow, D. E. *Metric Methods for Analyzing Partially Ranked Data*, volume 34 of *Lecture Notes in Statistics*. Springer, 1985.

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

Diaconis, Persi. *Group representations in probability and statistics*. Institute of Mathematical Statistics Lecture Notes - Monograph Series. Institute of Mathematical Statistics, Hayward, CA, 1988. ISBN 0-940600-14-5.

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

Fligner, M. A. and Verducci, J. S. Distance based ranking models. *JRSS Series B (Methodological)*, 48(3):359–369, 1986.

Guiver, John and Snelson, Edward. Bayesian inference for plackett-luce ranking models. In *ICML*, 2009.

Huang, J. and Guestrin, C. Riffled independence for ranked data. In *Proceedings of NIPS'09*, 2009.

Huang, J., Guestrin, C., and Guibas, L. Fourier theoretic probabilistic inference over permutations. *JMLR*, 10: 997–1070, 2009.

Huang, Jonathan, Kapoor, Ashish, and Guestrin, Carlos. Riffled independence for efficient inference with partial ranking. *Journal of Artificial Intelligence*, 44:491–532, 2012.

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

Hunter, David R. MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics*, 32:384–406, 2004.

Irurozki, Ekhine, Calvo, Borja, and Lozano, J. Learning probability distributions over permutations by means of Fourier coefficients. *Advances in Artificial Intelligence*, pp. 186–191, 2011.

Jagabathula, Srikanth and Shah, Devavrat. Inferring Rankings Using Constrained Sensing. *IEEE Transactions on Information Theory*, 57(11):7288–7306, 2011.

Kakarala, Ramakrishna. A signal processing approach to Fourier analysis of ranking data: the importance of phase. *IEEE Transactions on Signal Processing*, pp. 1–10, 2011.

Kakarala, Ramakrishna. Interpreting the phase spectrum in Fourier Analysis of partial ranking data. *Advances in Numerical Analysis*, 2012.

Kitaev, S. *Patterns in Permutations and Words*. Springer Publishing Company, Incorporated, 1st edition, 2011.

Kondor, Risi and Barbosa, Marconi S. Ranking with kernels in Fourier space. In *Proceedings of COLT'10*, pp. 451–463, 2010.

Kondor, Risi and Dempsey, Walter. Multiresolution analysis on the symmetric group. In *Neural Information Processing Systems 25*, 2012.

Lebanon, G. and Mao, Y. Non-parametric modeling of partially ranked data. *JMLR*, 9:2401–2429, 2008.

Lebanon, Guy and Lafferty, John. Cranking: Combining rankings using conditional probability models on permutations. In *Proceedings of the 19th International Conference on Machine Learning*, pp. 363–370, 2002.

Liqun, Xu. A multistage ranking model. *Psychometrika*, 65(2):217–231, 2000.

Lu, Tyler and Boutilier, Craig. Learning mallows models with pairwise preferences. In *ICML*, pp. 145–152, 2011.

Luce, R. D. *Individual Choice Behavior*. Wiley, 1959.

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

Meek, Christopher and Meila, Marina. Recursive inversion models for permutations. In *Advances in Neural Information Processing Systems 27*, pp. 631–639, 2014.

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

Rajkumar, Arun and Agarwal, Shivani. A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

Sun, Mingxuan, Lebanon, Guy, and Kidwell, Paul. Estimating probabilities in recommendation systems. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(3):471–492, 2012.

Wauthier, Fabian, Jordan, Michael, and Jojic, Nebojsa. Efficient ranking from pairwise comparisons. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 109–117, 2013.