Multiresolution analysis of incomplete rankings with applications to prediction

Eric Sibony
LTCI UMR No. 5141
Telecom ParisTech/CNRS
Institut Mines-Telecom
Paris, 75013, France
Email: eric.sibony@telecom-paristech.fr

Stéphan Clémençon
LTCI UMR No. 5141
Telecom ParisTech/CNRS
Institut Mines-Telecom
Paris, 75013, France

Jérémie Jakubowicz
SAMOVAR UMR No. 5157
Telecom SudParis/CNRS
Institut Mines-Telecom
Evry, 91000, France

***Abstract*—Data representing preferences of users are a typical example of the Big Datasets modern technologies, such as e-commerce portals, now permit to collect, in an explicit or implicit fashion. Such data are highly complex, insofar as the number of items n for which users may possibly express their preferences is explosive and the collection of items or products a given user actually examines and is capable of comparing is highly variable and of extremely low cardinality compared to n. It is the main purpose of this paper to promote a new representation of preference data, viewed as *incomplete rankings*. In contrast to alternative approaches, the very nature of preference data is preserved by the "multiscale analysis" we propose, identifying here "scale" with the set of items over which preferences are expressed, whose construction relies on recent results in algebraic topology. The representation of preference data it provides shares similarities with wavelet multiresolution analysis on a Euclidean space and can be computed at a reasonable cost given the complexity of the original data. Beyond computational and theoretical advantages, the "wavelet like" transform is shown to compress preference data into relatively few basis coefficients and thus facilitates statistical tasks such as distribution estimation or prediction. This is illustrated here by very encouraging empirical work based on popular benchmark real datasets.**

***Index Terms*—incomplete rankings, multiresolution analysis, preference data**

I. Introduction

With the ubiquity of sensors (*e.g.* mobile phones, internet, social networks) and the recent development of Big Data management technologies, the preferences of users regarding a wide variety of products such as movies, songs or books, restaurants, research papers or even financial services among others can be observed in real time. This brought the opportunity to design recommender engines predicting to which extent a user may prefer a given item compared to other items. Recommender systems have been increasingly popular these last few years, providing users with means to find their way in larger and larger digital catalogs and companies with the possibility to personalize their offers. Dedicated applications exploit historical data to produce personalized recommendation lists through *content-based* or *collaborative filtering* (or else through hybrid approaches), see [1] and the references therein. In this paper, focus is on the second approach, where the prediction of a list of items hopefully ranked according the preferences of a given user is based on those expressed by similar users. The present paper does not investigate the issue of defining an appropriate notion of similarity but focuses on a very different facet of the analysis of preference data. It tackles the challenging problem of representing efficiently the distribution of observed preference data. The major difficulty lies in the fact that such data are of the form of *incomplete rankings*: for each observation, only a few items are ranked and the collection of ranked items is highly variable. In this respect, none of the approaches proposed in the literature to process these data is fully satisfactory. Most of them either relies on the reduction of the observations to pairwise comparisons, see [2] for instance, or else consists in fitting a parametric probability distribution on the set of all permutations of the items and viewing the available data as truncated realizations of this distribution, see [3], [4] or more recently [5], [6]. Whereas the former approach does not exploit the whole information generally carried by preference data, the latter experiences difficulties in capturing the variability of real data, as documented in [7] for instance. Nonparametric kernel-based methods to model probability distributions on the symmetric group $\mathfrak{S}_n$, identified as the space of *complete rankings*, have also been recently considered in the literature, see [8] and [9], but do not scale to a large number of observations. In contrast, the framework we develop provides a representation which preserves the "multiscale" nature of data expressing preferences, "scale" being here assimilated to the collection of items actually ranked for each observation, see [10]. Although it relies on the topological properties of the *complex of injective words*, our approach to represent the distribution of preference data is directly inspired by multiresolution wavelet analysis on $\mathbb{R}^d$. These last two decades, novel harmonic analysis tools such as wavelet bases and their extensions have indeed revitalized signal and image processing and high-dimensional data analysis, leading to sparse representations and efficient algorithms for a wide variety of statistical tasks: estimation, prediction, denoising, compression, source separation, clustering, *etc*. In a similar manner, the concepts we consider in this paper to handle preference data can be used to solve a variety of statistical learning problems (*e.g.* efficient/sparse representation of rankings, ranking aggregation), paving the way for a novel approach to

collaborative filtering in particular. It is namely the main goal of the present article to propose a new method for predicting (incomplete) rankings based on preference data, fully relying on the notion of multiresolution analysis (MRA in abbreviated form) introduced in [10]. Beyond a precise description of the methodology we promote, experimental results based on the NETFLIX dataset (see [11]) are also reported, illustrating its relevance. Computational issues, related to scalability in particular, are also discussed in depth.

The rest of the paper is structured as follows. Section II introduces the main notations used throughout the paper and provides a rigorous formulation of the ranking prediction problem considered here. Optimal elements are exhibited in particular. Our approach to predictive ranking is next described at length in section III, where the specific notion of MRA on which the statistical method we promote is based is in particular briefly recalled. The implementation of the procedure introduced in the present paper together with scaling issues are discussed in depth in section IV. Preliminary experimental results based on real datasets are displayed in section V for illustration purpose.

II. BACKGROUND AND PRELIMINARIES

In this section, the general formalism and the main notations used throughout the paper are introduced. For any set E of finite cardinality $|E| < \infty$, we denote by $L(E) = \{f : E \to \mathbb{R}\}$ the set of real-valued functions on E and set $\mathcal{P}(E) = \{A \subset E \mid |A| \geq 2\}$. The indicator function of a subset $S \subset E$ is denoted by $\mathbb{1}_S$, while the indicator of any event $\mathcal{E}$ is denoted by $\mathbb{I}\{\mathcal{E}\}$, so that $\mathbb{1}_S(x) = \mathbb{I}\{x \in S\}$ for $x \in E$. The indicator function of a singleton $\{x\}$ is called a Dirac function, and denoted by δ_x.

A. Mathematical framework

We consider a statistical population of users expressing preferences as incomplete rankings on the set of items $[\![n]\!] = \{1, \ldots, n\}$ with $n \geq 1$. The data are thus of the form $(A_1, \pi^{(1)}), \ldots, (A_N, \pi^{(N)})$ where each A_i is a subset of $[\![n]\!]$ with at least two elements and each π_i is a ranking over A_i. Each ranking is assimilated to an *injective word*, *i.e.* an expression of the form $\pi = \pi_1 \ldots \pi_k$, where π_1 is the item ranked first, $\ldots$, π_k the item ranked last. The content of a ranking π is the set $c(\pi) = \{\pi_1, \ldots, \pi_k\}$ and its length is $|\pi| = |c(\pi)|$. For any $A \in \mathcal{P}([\![n]\!])$, we denote by $\Gamma(A)$ the set of all rankings of content A and by Γ_n the set of all incomplete rankings.

Equipped with these notations, a predictive model for incomplete rankings then takes the abstract form

$$\widehat{\pi} = h(A) \text{ with } h(A) \text{ in } \Gamma(A),$$

where the prediction rule maps a subset A of $\mathcal{P}([\![n]\!])$ to a ranking in $\Gamma(A) \subset \Gamma_n$. For this predictive task, the input space is $\mathcal{P}([\![n]\!])$, while Γ_n shall be referred to as the output space. The goal is to use historical data in order to produce an accurate mapping. Given the nature of the output values, an adequate notion of accuracy is defined by a loss function of the following type:

$$l : \bigcup_{A \in \mathcal{P}([\![n]\!])} \Gamma(A) \times \Gamma(A) \to \mathbb{R}_+.$$

Examples are provided in subsection II-B below. In the statistical learning framework we develop, the set of preference data on which a predictive rule is trained are modelled as $N \geq 1$ i.i.d. realizations $(A_1, \pi^{(1)}), \ldots, (A, \pi^{(N)})$ of a probability distribution μ on $\bigcup_{A \in \mathcal{P}([\![n]\!])} A \times \Gamma(A)$. Such a probability distribution is necessarily of the form

$$\mu(A, \pi) = \nu(A) p_A(\pi),$$

where $\nu(A)$ represents the probability of observing the preference of a user in the statistical population of interest on the subset of items A, and $p_A(\pi)$ represents the probability of observing the specific ranking π on A, conditioned upon the observation of a ranking on A. The theoretical risk for the problem of incomplete rankings prediction is therefore defined, for a classifier h, as the expectation

$$\mathcal{R}(h) = \sum_{A \in \mathcal{P}([\![n]\!])} \nu(A) \sum_{\pi \in \Gamma(A)} l(h(A), \pi) p_A(\pi), \quad (1)$$

and the related empirical risk by

$$\mathcal{R}_N(h) = \sum_{i=1}^{N} l(h(A_i), \pi_i). \quad (2)$$

B. Optimality

The following result exhibits an optimal classifier regarding the (theoretical) risk minimization problem stated previously. Its proof is straightforward and left to the reader.

Proposition 1 (Optimal classifier)**.** *Let h^* be any predictive rule such that $h^*(A)$ is a solution of the minimization problem*

$$\min_{\pi \in \Gamma(A)} \sum_{\pi' \in \Gamma(A)} l(\pi, \pi') p_A(\pi') \quad (3)$$

for all $A \in \mathcal{P}([\![n]\!])$. Then, the predictive rule h^* has minimum risk (1)*.*

Notice first that an optimal prediction rule h^* always exists: the set $\Gamma(A)$ is of finite cardinality and, thus, there always exists a solution to the minimization problem (3). It is however not necessarily unique. From a practical perspective, observe also that such a solution is not accessible, because the p_A's are unknown. If the restriction of the loss function to $\Gamma(A)^2$, $l_{|\Gamma(A)^2}$, is a metric d_A on $\Gamma(A)$, then $h^*(A)$ can be viewed as a "consensus" ranking for the distribution p_A with respect to d_A. There is a wide literature about the computation of exact or approximate consensus rankings, especially for the Kendall tau distance, defined as the number of discordant pairwise comparisons:

$$d_A(\pi, \pi') = |\{1 \leq i < j \leq |A| \mid (\pi_j - \pi_i)(\pi'_j - \pi'_i) < 0\}|,$$

for $\pi, \pi' \in \Gamma(A)$, see [12]. In the case of the $0-1$ loss function $l(\pi, \pi') = \mathbb{I}\{\pi \neq \pi'\}$, problem (3) becomes

$$\min_{\pi_0 \in \Gamma(A)} \sum_{\pi \neq \pi_0} p_A(\pi) = \min_{\pi_0 \in \Gamma(A)} (1 - p_A(\pi_0))$$

and the optimal classifier h^* is then defined by $h^*(A) = \mathrm{argmax}_{\pi \in \Gamma(A)} p_A(\pi)$. In this paper, we do not specify any loss function but focus on the estimation of the p_A's based on the available data instead, which then allows to construct an approximate version of h^* according to the *plug-in* paradigm, see [13] or [14].

C. The statistical nature of the predictive problem

For $A \in \mathcal{P}(\llbracket n \rrbracket)$ and $\pi \in \Gamma(A)$, $p_A(\pi)$ represents the chance of observing the ranking π on the items of A among all the possible rankings on A. If there were no relation between the p_A's, the problem of predicting incomplete rankings would then boil down to $2^n - n - 1$ independent problems, consisting each in predicting a full ranking on a fixed subset $A \in \mathcal{P}(\llbracket n \rrbracket)$ of items. Yet it seems reasonable to expect that, in most situations encountered in practice, the distribution of the preferences regarding the pair of items $\{a, b\}$ carries information about the distribution of the preferences regarding the subset of items $\{a, b, c\}$. In the literature dedicated to ranking, is common to assume the existence of a probability distribution p on $\mathfrak{S}_n$ such that for all $A \in \mathcal{P}(\llbracket n \rrbracket)$ and $\pi \in \Gamma(A)$,

$$p_A(\pi) = \sum_{\sigma \in \mathfrak{S}_n(\pi)} p(\sigma), \tag{4}$$

where $\mathfrak{S}_n(\pi) = \{\sigma \in \mathfrak{S}_n \mid \exists 1 \leq i_1 < \cdots < i_{|\pi|} \leq n \text{ s.t. } \pi = \sigma_{i_1} \dots \sigma_{i_{|\pi|}}\}$ is the set of the linear extensions of π to $\llbracket n \rrbracket$. The probability distribution p is then referred to as the *ranking model* and p_A as the marginal of p on A. More generally, we define the *marginal operator* $M_A : L(\mathfrak{S}_n) \to L(\Gamma(A))$ on $A \in \mathcal{P}(\llbracket n \rrbracket)$ by

$$M_A f(\pi) = \sum_{\sigma \in \mathfrak{S}_n(\pi)} f(\sigma)$$

for $f \in L(\mathfrak{S}_n)$ and $\pi \in \Gamma(A)$. It is generally vain to attempt to estimate p in practice in absence of any additional structural assumption. Indeed available observations usually describe preferences over subsets $A \in \mathcal{P}(\llbracket n \rrbracket)$ of very low cardinality compared to n, providing a censored information solely. However, from a predictive ranking perspective, predictions need to be evaluated on observable subsets only. In the subsequent analysis, we do not consider the problem of estimating the distribution ν and take it equal to its empirical estimator $\widehat{\nu}_N(A) = |\{1 \leq i \leq N \mid A_i = A\}|/N$. Predictions are then evaluated on its support $\mathcal{A} = \{A \in \mathcal{P}(\llbracket n \rrbracket) \mid \nu(A) > 0\}$, which shall be referred to as the *observation design*. Now, it is nevertheless useful in practice to be able to predict rankings on any subset $A \in \mathcal{P}(\llbracket n \rrbracket)$, and this is naturally achieved in a consistent fashion by constructing a ranking model and then computing its marginals. Notice however that it is not mandatory to impose the non-negativity of the ranking model for the purpose of ranking prediction, the definition (3) of optimal classifiers still being valid. Hence the following problem.

Problem 1. *Find a function $\tilde{p} \in L(\mathfrak{S}_n)$ such that $M_A \tilde{p} = p_A$ for all $A \in \mathcal{A}$.*

If the p_A's are known, this problem boils down to solving a linear system, of disarming simplicity at first glance, but actually presenting a great computational challenge. A ranking model $\tilde{p}$ is indeed described by parameter of dimension $n! - 1$ and for $|A| = k$, the naive computation of $M_A \tilde{p}(\pi)$ requires to sum over $n!/k!$ terms for each $\pi \in \Gamma(A)$. If n is around 10^4 and k around 5, this is by far intractable. Hopefully, the framework recently developed in [10] to define a multiresolution analysis (MRA) of incomplete rankings allows to cope with this system by expressing it very parsimoniously in an appropriate "wavelet basis". We point out that in practice the p_A's are unknown, so that the following statistical version of Problem 1 should be considered.

Problem 2. *Based on observations $(A_1, \pi_1), \dots, (A_N, \pi_N)$, find a function $\tilde{p}_N \in L(\mathfrak{S}_n)$ such that $\mathbb{E}[M_A \tilde{p}_N] = p_A$ for all $A \in \mathcal{A}$.*

In the next section it is shown at length how to use the MRA framework for incomplete rankings introduced in [10] in order to provide computationally feasible solutions to both of these problems.

D. Related work - Competitors

As recalled in the Introduction section, though of limited accuracy due to too rigid structural assumptions, statistical methods based on the Luce-Plackett model (see [3], [4]) can be extended to incomplete rankings and scale without difficulty. The *Luce's choice axiom* permits to drastically reduce the model complexity, encapsulated by n parameters solely (in contrast with the cardinality $n!$ of $\mathfrak{S}_n$). It has been used in a wide variety of applications and several algorithms have been proposed to infer its parameters, see [5], [6] or [15] for instance. Several numerical experiments on real datasets have shown however that its capacity to fit real data is very limited, the model being too rigid to handle the singularities observed in practice, refer to [7] and [9]. Two alternative nonparametric kernel-based approaches have been proposed. A diffusion kernel is used in the Fourier domain in [8], whereas a triangular kernel with respect to the Kendall's tau distance is considered in [9]. Both approaches deal however with sets $\mathfrak{S}_n(\pi)$ and not directly with incomplete rankings π. This tends to blend the estimated probabilities of occurence of the incomplete rankings, inducing a statistical bias as well as a higher computational complexity. In contrast, the framework we develop relies on the natural multiresolution structure of incomplete rankings. To the best of our knowledge, it is the first that permits to define approximation procedures directly on the ranked data.

We also point out that, in [16], a multiresolution analysis on $L(\mathfrak{S}_n)$ has been proposed. It relies on a multiscale structure for $\mathfrak{S}_n$ based on the embedding of subgroups $\mathfrak{S}_1 \subset \ldots \subset$

$\mathfrak{S}_{n-1} \subset \mathfrak{S}_n$ given by $\mathfrak{S}_k \simeq \{\sigma \in \mathfrak{S}_n \mid \sigma(n-j) = n-j \quad \text{for } j = 0, ..., k-1\}$. It is a first breakthrough in dealing with singularities of probability distributions on rankings. However, singularities corresponding to incomplete rankings cannot be localized with this construction, because incomplete rankings do not interact well with the group structure of $\mathfrak{S}_n$. Several approaches have been proposed to generalize the construction of multiresolution analyses and wavelet bases on discrete spaces, mostly on trees and graphs, see for instance [17], [18], [19], [20] and [21]. None of them leads however to our construction for incomplete rankings, which crucially relies on the topological properties of the complex of injective words.

The use of topological tools to analyze ranked data has been introduced in [22] and then pursued in several contributions such as in [23] or [24]. Their approach consists in modeling a collection of pairwise comparisons as an oriented flow on the graph with vertices $[\![n]\!]$ where two items are linked if the pair appears at least once amo,g the comparisons observed. They show that this flow admits a "Hodge decomposition" in the sense that it can be decomposed as the sum of three components, a "gradient flow" that corresponds to globally consistent rankings, a "curl flow" that corresponds to locally inconsistent rankings, and a "harmonic flow", that corresponds to globally inconsistent but locally consistent rankings. Our construction also relies on results in topology but it decomposes the information in a very different manner, tailored to the situation where incomplete rankings of any size may possibly be observed.

III. MRA BASED STATISTICAL ESTIMATION

It is the main purpose of this section to recall key notions about MRA of incomplete rankings and to show how to it can be used to define a ranking model. By convention, we define the empty word $\overline{0}$ as the unique word of content $\emptyset$ and length 0.

A. Definitions and properties

For a word $\pi \in \Gamma_n$ and $A \subset c(\pi)$, we denote by $\pi_{|A}$ the subword of π obtained by keeping the letters in A only. It represents the ranking induced by π over A. By definition, $M_A \delta_\sigma = \delta_{\sigma_{|A}}$ for $A \in \mathcal{P}([\![n]\!])$ and $\sigma \in \mathfrak{S}_n$. We extend the marginal operator to $L(\Gamma_n)$, still denoting it by M_A, by setting for any $\pi \in \Gamma_n$

$$M_A \delta_\pi = \begin{cases} \delta_{\pi_{|A}} & \text{if } A \subset c(\pi), \\ 0 & \text{otherwise.} \end{cases}$$

The fundamental result of multiresolution analysis of incomplete rankings is that for any $A \in \mathcal{P}([\![n]\!])$, any function $f \in L(\Gamma(A))$ can be decomposed as a sum of components that each localize the information specific to the marginal of f on a subset $B \in \mathcal{P}(A)$. Let us introduce some more notations to be more specific. For $\tau \in \mathfrak{S}_n$, we define $\mathrm{supp}(\tau) = \{1 \leq i \leq n \mid \tau(i) \neq i\}$. In [10], it is shown how to build a family $(x_\tau)_{\tau \in \mathfrak{S}_n}$ of functions on Γ_n together with a family of embedding operators $(\phi_A)_{A \in \mathcal{P}([\![n]\!])}$, $\phi_A : L(\Gamma_n) \to L(\Gamma(A))$, so that for all $A \in \mathcal{P}([\![n]\!])$, $(\phi_A x_\tau)_{\mathrm{supp}(\tau) \subset A}$ is a basis of $L(\Gamma(A))$ (the operator $\phi_{[\![n]\!]}$ is abusively denoted by ϕ_n by convention).

Remark 1. The function x_{id} is actually not defined originally in [10], we define it here by $x_{id} = \delta_{\overline{0}}$. We also extend the ϕ_A's by setting $\phi_A x_{id} = \mathbb{1}_{\Gamma(A)}$.

This basis can be interpreted as a "wavelet basis" because the coefficients in the expansion of a function each localize a specific piece of information about its marginals. Indeed, let $\overline{\phi}_A$ be the following normalized version of the embedding operator ϕ_A

$$\overline{\phi}_A \delta_\pi = \begin{cases} \dfrac{1}{|A|!} \phi_A \delta_\pi & \text{if } \pi = \overline{0}, \\ \dfrac{1}{(|A| - |\pi| + 1)!} \phi_A \delta_\pi & \text{otherwise,} \end{cases}$$

for $A \in \mathcal{P}([\![n]\!])$. For $f \in L(\Gamma(A))$, let $(c_\tau(f))_{\mathrm{supp}(\tau) \subset A}$ be the coefficients of its expansion in the basis $(\overline{\phi}_A x_\tau)_{\mathrm{supp}(\tau) \subset A}$, *i.e.* such that

$$f = \sum_{\mathrm{supp}(\tau) \subset A} c_\tau(f) \overline{\phi}_A x_\tau.$$

Then by virtue of Proposition 10 of [10], we have:

$$M_B f = \sum_{\mathrm{supp}(\tau) \subset B} c_\tau(f) \overline{\phi}_B x_\tau \tag{5}$$

for any $B \in \mathcal{P}(A)$. Hence, only the coefficients $c_\tau(f)$ for $\mathrm{supp}(\tau) \subset B$ are involved in the computation of the marginal of f over B. By substraction, this means that the piece of information carried by the coefficients $c_\tau(f)$ for $\mathrm{supp}(\tau) = B$ is exactly that which is specific to the marginal of f over B. Note that formula (5) also implies that

$$c_\tau\left(M_B f\right) = c_\tau(f) \tag{6}$$

for $\mathrm{supp}(\tau) \subset B$.

Example 1. Figure 1 exhibits the normalized wavelet basis $(\overline{\phi}_3 x_\tau)_{\tau \in \mathfrak{S}_3}$ of $L(\Gamma(\{1,2,3\}))$. Each function is represented on a graph with x-axis representing the elements of $\Gamma(\{1,2,3\})$ in the lexicographic order $123, 132, \ldots, 321$. There is one wavelet of order 0, namely the constant function $\overline{\phi}_3 x_{id}$, three wavelets of order 2, one for each subset of size 2 in $\{1,2,3\}$, and two wavelets of order 3, the latter localizing the piece of information specific to the marginal on $\{1,2,3\}$. More generally, for a subset of items $A \in \mathcal{P}([\![n]\!])$, the piece of information specific to A is localized by a number of $\#\{\tau \in \mathfrak{S}_n \mid \mathrm{supp}(\tau) = A\}$ basis elements, equal to the number of fixed-point free permutations on a set with $|A|$ elements.

By its powerful localization properties, multiresolution analysis of incomplete rankings provides a simple way to solve both Problems 1 and 2.

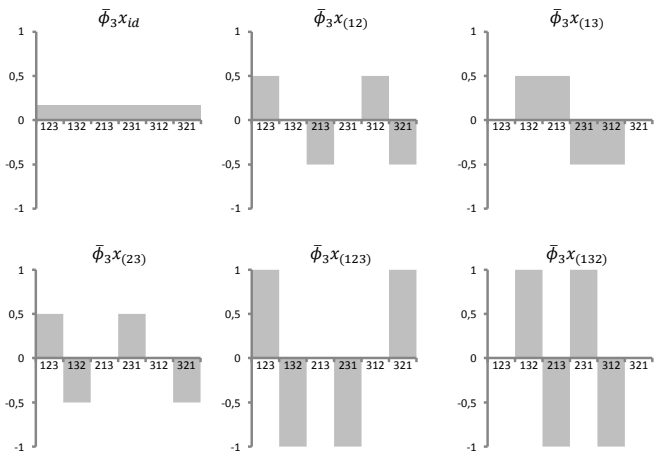


Fig. 1. Normalized wavelet basis of $L(\Gamma(\{1,2,3\}))$

B. Solutions to Problems 1 and 2

Formula (5) shows that Problem 1 boils down to finding a function $\tilde{p} \in L(\mathfrak{S}_n)$ such that $c_\tau(\tilde{p}) = c_\tau(p_A)$ for all $\tau \in \mathfrak{S}_n$ satisfying the condition $\mathrm{supp}(\tau) \in \bigcup_{A\in\mathcal{A}} \mathcal{P}(A)$. Let $\mathfrak{S}_\mathcal{A}$ denote the set of such τ's:

$$\mathfrak{S}_\mathcal{A} := \{\tau \in \mathfrak{S}_n \mid \mathrm{supp}(\tau) \in \bigcup_{A\in\mathcal{A}} \mathcal{P}(A)\}.$$

Now, recalling that $p_A = M_A p$ for $A \in \mathcal{A}$ by definition, Eq. (6) implies that for $\tau \in \mathfrak{S}_n$, $c_\tau(p_A) = c_\tau(p)$ for all $A \in \mathcal{A}$ such that $\mathrm{supp}(\tau) \subset A$. This means that in the theoretical setting of Problem 1 where the p_A's are considered to be known, one directly has access to the coefficients $c_\tau(p)$ for all $\tau \in \mathfrak{S}_\mathcal{A}$, leading to a natural solution for Problem 1. This is summarized in the following theorem.

Theorem 1. *If all the p_A's are known then the $c_\tau(p)$'s are known for all $\tau \in \mathfrak{S}_\mathcal{A}$, and the function $\tilde{p}$ defined by*

$$\tilde{p} = \sum_{\tau\in\mathfrak{S}_\mathcal{A}} c_\tau(p)\overline{\phi}_n x_\tau$$

is a solution to Problem 1.

In the statistical setting the p_A's are not considered as known any more and Problem 2 boils down to finding unbiased estimators of the $c_\tau(p)$'s for all $\tau \in \mathfrak{S}_\mathcal{A}$. We thus consider the empirical estimators, defined for $A \in \mathcal{P}(\llbracket n \rrbracket)$ by

$$\widehat{p}_A(\pi) = \frac{|\{1 \le i \le N \mid \pi_i = \pi\}|}{|\{1 \le i \le N \mid A_i = A\}|}.$$

Since $\mathbb{E}[\widehat{p}_A(\pi)] = p_A(\pi)$ for all $\pi \in \Gamma(A)$, one has $\mathbb{E}[c_\tau(\widehat{p}_A)] = c_\tau(p_A)$. For a fixed $\tau \in \mathfrak{S}_\mathcal{A}$, $c_\tau(\widehat{p}_A)$ is then an unbiased estimator of $c_\tau(p)$ for any $A \in \mathcal{A}$ such that $\mathrm{supp}(\tau) \subset A$. However, each of these estimators may have a large variance, therefore we average them so as to produce the estimator with reduced variance

$$\tilde{c}_\tau = \sum_{\substack{A\in\mathcal{A}\\ \mathrm{supp}(\tau)\subset A}} \nu(A) c_\tau(\widehat{p_A}). \tag{7}$$

The corresponding function $\tilde{p}_N$ is then defined by

$$\tilde{p}_N = \sum_{\tau\in\mathfrak{S}_\mathcal{A}} \tilde{c}_\tau \overline{\phi}_n x_\tau, \tag{8}$$

and the result stated below is straightforward.

Theorem 2. *The function $\tilde{p}_N$ defined by Eq.* (8) *is a solution to Problem 2.*

The function $\tilde{p}_N$ is the estimator we propose to predict incomplete rankings in practice. Though its definition is intuitive once the MRA framework is set up, we provide more insight by exhibiting a second interpretation, in terms of weighted least squares estimation (LSE in short).

C. A weighted LSE in the feature space

By a simple sum inversion, the estimator $\tilde{p}_N$ defined in (8) can be rewritten as

$$\tilde{p}_N = \sum_{A\in\mathcal{A}} \nu(A) \sum_{\mathrm{supp}(\tau)\subset A} c_\tau(\widehat{p}_A)\overline{\phi}_n x_\tau. \tag{9}$$

This formula can be interpreted as follows. For each $A \in \mathcal{P}(\llbracket n \rrbracket)$, $\widehat{p}_A$ is a function in $L(\Gamma_n)$. The "wavelet transform" $\Psi : \widehat{p}_A \mapsto (c_\tau(\widehat{p}_A))_{\mathrm{supp}(\tau)\subset A}$ can be viewed as a mapping from the "signal space" $L(\Gamma_n)$ to the "feature space" $\mathbb{R}^{n!}$, by extending the collection of coefficients $(c_\tau(\widehat{p}_A))_{\mathrm{supp}(\tau)\subset A}$ to $(c_\tau(\widehat{p}_A))_{\tau\in\mathfrak{S}_n}$ where $c_\tau(\widehat{p}_A) = 0$ if $\mathrm{supp}(\tau) \not\subset A$. The wavelet transform of $\tilde{p}_N$ is thus computed as the average of the wavelet transforms of the $\widehat{p}_A$'s, each being weighted by its frequency of occurrence $\nu(A)$. In other words, $\tilde{p_N}$ is the solution to the following minimization problem

$$\min_{q\in L(\mathfrak{S}_n)} \sum_{A\in\mathcal{A}} \nu(A) \|\Psi(q) - \Psi(\widehat{p_A})\|_2^2, \tag{10}$$

where $\|\cdot\|_2$ is the usual Euclidean norm on $\mathbb{R}^{n!}$ and therefore

$$\|\Psi(f)\|_2^2 = \sum_{\tau\in\mathfrak{S}_n} c_\tau(f)^2$$

for any $f \in L(\Gamma_n)$.

Remark 2. The wavelet basis constructed in [10] for $L(\mathfrak{S}_n)$ is not orthogonal and thus for $f \in L(\mathfrak{S}_n)$, we generally have $\sum_{\sigma\in\mathfrak{S}_n} f(\sigma)^2 \neq \sum_{\tau\in\mathfrak{S}_n} c_\tau(f)^2$.

D. How to calculate the wavelet coefficients

We now explain how to compute the wavelet transform $f \mapsto \Psi(f) = (c_\tau(f))_{\tau\in\mathfrak{S}_n}$. For $A \in \mathcal{P}(\llbracket n \rrbracket)$ and $f \in L(\Gamma(A))$, the $c_\tau(f)$'s for $\mathrm{supp}(\tau) \subset A$ are defined as the coefficients of the expansion of f in the basis $(\overline{\phi}_A x_\tau)_{\mathrm{supp}(\tau)\subset A}$. They can thus be obtained by inverting the linear system defined by the $|A|!$ equations

$$\begin{cases} \displaystyle\sum_{\mathrm{supp}(\tau)\subset A} \overline{\phi}_A x_\tau(\pi) c_\tau(f) = f(\pi) \\ \qquad\qquad\qquad\qquad \vdots \end{cases}$$

for $\pi \in \Gamma(A)$. Performing this computation for each new observation or even precomputing the inversions for all $A \in \mathcal{A}$ is

however intractable in practice. Hopefully, two results allow to reduce it drastically. First observe that for $B \in \mathcal{P}(A)$, $c_\tau(f) = c_\tau(M_B f)$ for all $\tau \in \mathfrak{S}_n$ such that $\mathrm{supp}(\tau) = B$, by equation (6). This implies that the coefficients $(c_\tau(f))_{\mathrm{supp}(\tau)=B}$ can be computed by inverting the smaller system of $|B|!$ equations

$$\begin{cases} \displaystyle\sum_{\mathrm{supp}(\tau)=B} x_\tau(\pi) c_\tau(f) = M_B f(\pi) \\ \qquad\qquad\qquad \vdots \end{cases}$$

for $\pi \in \Gamma(B)$. In other words, the computation of $c_\tau(f)$ only requires to know the values of the marginal of f on $\mathrm{supp}(\tau)$ instead of all the values taken by f. For $\tau \in \mathfrak{S}_n$, we denote by $(\alpha_\tau(\pi))_{\pi\in\Gamma(\mathrm{supp}(\tau))}$ the scalars such that

$$c_\tau(f) = \sum_{\pi\in\Gamma(\mathrm{supp}(\tau))} \alpha_\tau(\pi) M_{\mathrm{supp}(\tau)} f(\pi) \tag{11}$$

for any $f \in L(\Gamma_n$ or equivalently defined by

$$\alpha_\tau(\pi) = c_\tau(\delta_\pi) \tag{12}$$

for all $\pi \in \Gamma(\mathrm{supp}(\tau))$.

The second result stems from the construction of the wavelet basis. Proposition 8 of [10] indeed states that for any $\tau \in \mathfrak{S}_n$, the wavelet function x_τ is obtained from a wavelet $x_{\tau'}$ with $\mathrm{supp}(\tau') = \{1, \dots, |\mathrm{supp}(\tau)|\}$ by a simple relabeling. This directly implies that α_τ is obtained by $\alpha_{\tau'}$ by the same relabeling. For instance, let $\tau = (245)$ and $\sigma : \{2,4,5\} \to \{1,2,3\}$ defined by $\sigma(2) = 1, \sigma(4) = 2$ and $\sigma(5) = 3$. Then for $\pi \in \Gamma(\{2,4,5\})$, $x_{(245)}(\pi) = x_{(123)}(\sigma(\pi))$ and $\alpha_{(245)}(\pi) = \alpha_{(123)}(\sigma(\pi))$, where $\sigma(\pi)$ is the word $\sigma(\pi_1)\sigma(\pi_2)\sigma(\pi_3)$. This last fact means that to compute the wavelet transform of any function $f \in L(\Gamma(A))$ for $A \subset [\![n]\!]$ with $2 \le |A| \le k$, we only need to know how to compute the vectors of coefficients $(\alpha_\tau(\pi))_{\pi\in\Gamma(\mathrm{supp}(\tau))}$ for $\tau \in \mathfrak{S}_n$ with $\mathrm{supp}(\tau) = \{1, \dots, j\}$, for $j \in \{2, \dots, k\}$. In practice we calculate the exact formulas and hard-code them. For instance for $k = 2$, $\alpha_{(ab)}(ab) = 1/2$ and $\alpha_{(ab)}(ba) = -1/2$, *i.e.*

$$c_{(ab)}(f) = \frac{1}{2}\left(M_{\{a,b\}} f(ab) - M_{\{a,b\}} f(ba)\right).$$

Eventually, the following proposition shows how to calculate the coefficients $\tilde{c}_\tau$ involved in the definition of the estimator $\tilde{p}_N$. We shall use it to define map/reduce procedures in section IV. For $\pi, \pi' \in \Gamma_n$, we denote by $\pi \subset \pi'$ to say that π is a subword of π' (for instance $123 \subset 51243$, while $312 \not\subset 51243$).

Proposition 2. *For $\tau \in \mathfrak{S}_n$, the coefficient $\tilde{c}_\tau$ defined in* (7) *satisfies*

$$\tilde{c}_\tau = \frac{1}{N} \sum_{\pi\in\Gamma(\mathrm{supp}(\tau))} |\{1 \le i \le N \mid \pi \subset \pi_i\}| \alpha_\tau(\pi),$$

Proof. Inserting the expressions of $\widehat{p_A}(\pi)$ and $\nu(A)$ in (7), one obtains

$$\begin{aligned} \tilde{c}_\tau &= \frac{1}{N} \sum_{\substack{A\in\mathcal{A} \\ \mathrm{supp}(\tau)\subset A}} c_\tau\left(\sum_{\pi\in\Gamma(A)} |\{1 \le i \le N \mid \pi_i = \pi\}| \delta_\pi\right) \\ &= \frac{1}{N} \sum_{i=1}^{N} c_\tau(\delta_{\pi_i}), \end{aligned}$$

where we recall that $c_\tau(\delta_\pi) = 0$ if $\mathrm{supp}(\tau) \not\subset c(\pi)$ by convention. The proof is concluded using equation (11). □

E. Estimation and regularization

Once the estimator $\tilde{p}_N$ is calculated, it is used to estimate the probabilities of the rankings on any subset $A \in \mathcal{P}([\![n]\!])$ via formula (5), which gives

$$M_A \tilde{p}_N = \sum_{\mathrm{supp}(\tau)\subset A} \tilde{c}_\tau \overline{\phi}_A x_\tau. \tag{13}$$

We give more insight about this estimation procedure distinguishing two cases.

- A is a subset of an observed set, *i.e.* $A \in \bigcup_{B\in\mathcal{A}} \mathcal{P}(B)$. All the coefficients $c_\tau(p)$ for $\mathrm{supp}(\tau) \subset A$ are estimated, and $M_A \tilde{p}_N$ is an unbiased estimator of $M_A p$.
- A is not a subset of an observed set, *i.e.* $A \notin \bigcup_{B\in\mathcal{A}} \mathcal{P}(B)$. In this case, only a portion of the coefficients $c_\tau(p)$ for $\mathrm{supp}(\tau) \subset A$ are estimated, namely only those that also belong to $\mathfrak{S}_\mathcal{A}$. The estimator $M_A \tilde{p}_N$ is then biased

  $$\mathbb{E}\left[M_A \tilde{p}_N\right] = \sum_{\substack{\mathrm{supp}(\tau)\subset A \\ \tau\in\mathfrak{S}_\mathcal{A}}} c_\tau(p) \overline{\phi}_A x_\tau$$

  and can be seen as the regularized version where the non-observed coefficients have been put equal to 0, following the sparsity inducing paradigm. Other regularization procedures can be considered, they are left for future work.

IV. Implementation and scaling issues

From observation to prediction, the method introduced in this paper can be decomposed into three phases.

1) **Training phase:** the model is trained on the observations $(A_1, \pi_1), \dots, (A_N, \pi_N)$ and stored.
2) **Probability estimation:** for a subset of items $A \in \mathcal{P}([\![n]\!])$, the estimated probability distribution $M_A \tilde{p}_N$ is computed.
3) **Prediction:** a ranking on A is proposed, based on $M_A \tilde{p}_N$.

Phase 3 strongly depends on the choice of the loss function. Investigating how to perform it in a nearly optimal fashion in a fully general framework is beyond the scope of the present paper.

A. Training phase

From a formal perspective, the method we have introduced mainly consists in computing and storing the estimator $\tilde{p}_N$. Representing the latter as a distribution over $\mathfrak{S}_n$ is however unfeasible in practice because this would imply the computation and storage of the $n!$ values $\tilde{p}_N(\sigma)$ for $\sigma \in \mathfrak{S}_n$, which is intractable as soon as $n > 15$, whereas n is around 10^4 in many applications! By its powerful localization properties, MRA of incomplete rankings allows to overcome this difficulty. The estimator $\tilde{p}_N$ is indeed fully characterized by the coefficients $(\tilde{c}_\tau)_{\tau \in \mathfrak{S}_\mathcal{A}}$ defined in (7), and all its marginal projections as well, by Eq. (13). The training phase thus boils down to the computation and storage of these coefficients and the number of such coefficients satisfies the inequality

$$|\mathfrak{S}_\mathcal{A}| = \left| \bigcup_{A \in \mathcal{A}} \mathcal{P}(A) \right| \leq \sum_{A \in \mathcal{A}} 2^{|A|} \leq N 2^{k_0},$$

where N is the number of observations and k_0 is the maximum size of a possibly observed incomplete ranking. In our experiment on a dataset constructed from the NETFLIX dataset (see section V), $N \simeq 1.6 \times 10^8$ and $k_0 = 5$, leading to approximately 5×10^9 values. The number of coefficients is ridiculously small compared to $n!$ ($n \simeq 20000$ for the NETFLIX dataset), but still represents a large volume of complex data. The computation and the storage of the coefficients $(\tilde{c}_\tau)_{\tau \in \mathfrak{S}_\mathcal{A}}$ should be thus performed in a distributed key/value framework. The $\tilde{c}_\tau$'s are encoded as key/value pairs $\langle \texttt{tau}, \texttt{c} \rangle$, where `tau` identifies the associated wavelet and `c` is the corresponding coefficient, and their computation is distributed, based on Proposition 2. The dataflow for the training phase then consists in the following steps.

1) Collect the preferences encoded as words (for instance, the word `51243` means that object `5` is preferred to object `1`, itself preferred to object `2`, etc.) After this step is completed, we are left with a possibly very large list of words, stored as a distributed text file.
2) Count the number of times each word occurs as a subword of a preference in the file. This is a standard map/reduce job. After this step is completed, we are left with a distributed key/value database where each entry has the form $\langle \texttt{pi}, \texttt{n} \rangle$ where `pi` denotes the preference encoded as a word and `n` is an integer representing how many times `pi` was present as a subword of a word in the initial word list.
3) Perform the wavelet analysis *per se* as the map/reduce job illustrated by figure 2 (`alpha(tau,pi)` corresponds to precomputed $\alpha_\tau(\pi)$). After this step is completed the coefficients are stored as key/value pairs $\langle \texttt{tau}, \texttt{c} \rangle$ where `tau` identifies the associated wavelet and `c` is the corresponding coefficient.

The global workflow for the training phase is summarized in Fig. IV-A.

$$\begin{cases} \text{map}: & \langle \texttt{pi}, \texttt{n} \rangle \to (\langle \texttt{tau}, \texttt{alpha(tau,pi)} * \texttt{n} \rangle)_{tau \subset pi} \\ \text{reduce}: & (\langle \texttt{tau}, \ \texttt{c}_i \rangle)_{i \in I} \to \langle \texttt{tau}, \sum_{i \in I} \texttt{c}_i / N \rangle \end{cases}$$

Fig. 2. Wavelet analysis as a single map/reduce job

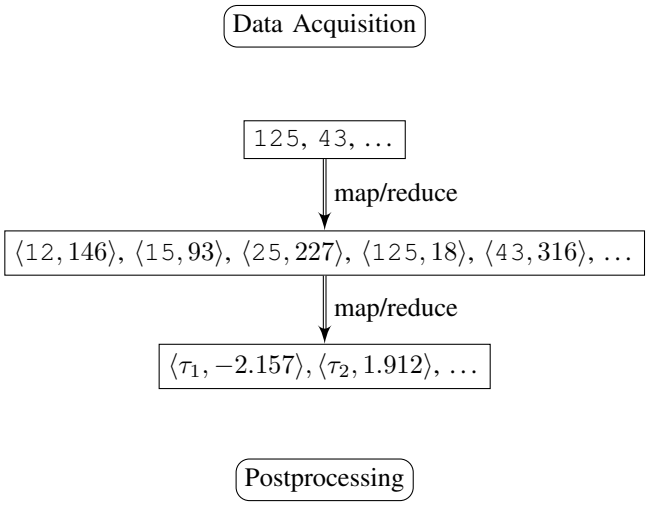


Fig. 3. Workflow for the training phase: acquire data, turn data into a key/value database, compute wavelet coefficients as a map/reduce job

B. Probability estimation

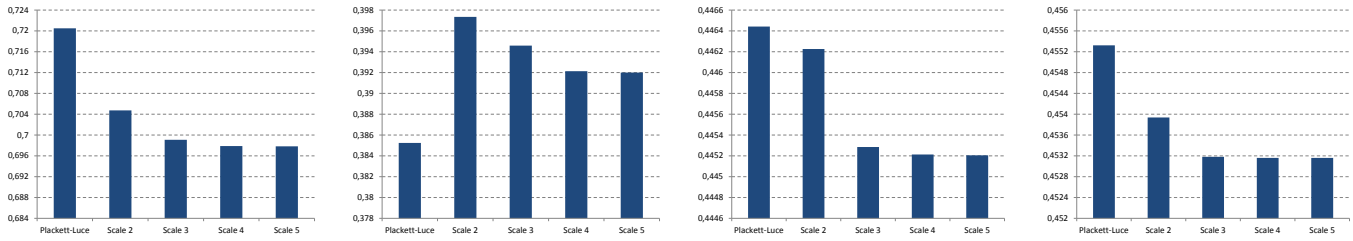
Once the coefficients $(\tilde{c}_\tau)_{\tau \in \mathfrak{S}_\mathcal{A}}$ are computed and stored, the estimation of the probabilities of the rankings on any subset $A \in \mathcal{P}([\![n]\!])$ is performed using formula (13). Thus for $|A| = k$, the computational complexity of $M_A \tilde{p}_N$ only depends on k. An analysis of this complexity will be done in future work.

V. NUMERICAL EXPERIMENTS

For illustration purpose, we present the results of numerical experiments conducted on two datasets, the SUSHI dataset ($n = 10$) and the NETFLIX dataset ($n = 17,770$). In both cases, we generate from raw data incomplete rankings of size 2 to 5 (the maximum size 5 is a consequence of the rating scale in the NETFLIX dataset).

A. Evaluation setting

The predictions are evaluated through four different loss functions: the $0-1$ loss, the Kendall's tau distance, the $l1$ distance (also called Spearman's footrule metric) and the $l2$ distance (also called Spearman's rho metric). The two latter are respectively defined, for any $A \in \mathcal{P}([\![n]\!])$ and $\pi, \pi' \in A$, by $d_1(\pi, \pi') = \sum_{a \in A} |\pi(a) - \pi'(a)|$ and $d_1(\pi, \pi') = \sum_{a \in A} (\pi(a) - \pi'(a))^2$, where for $a \in A$, $\pi(a)$ denotes the rank of the item a in π. All these distances are of course invariant under relabeling of the items, and can thus evaluate the accuracy of the predictions on different subsets of items of the same size in a consistent manner. In order to be fully consistent when dealing with subsets of different sizes, we use their normalized versions defined for $\pi, \pi' \in \Gamma(A)$



(a) $0-1$ Loss (b) Kendall's tau distance (c) $L1$ distance (d) $L2$ distance

Fig. 4. Empirical risk of the plug-in predictors on the SUSHI dataset

with $A \in \mathcal{P}(\llbracket n \rrbracket)$ by

$$\overline{d}(\pi, \pi') = \frac{d(\pi, \pi')}{\max_{(\omega,\omega') \in \Gamma(\{1,\dots,|A|\})^2} d(\omega, \omega')},$$

where d denotes any of the four distances. That way, each loss function takes its values in $[0, 1]$ for a couple of incomplete rankings of any size. A classifier is evaluated by its empirical risk on a test dataset with respect to a loss function. As a baseline, we compute the expectation and standard deviation of the risk of the uniformly random classifier that predicts, for any subset of items $A \in \mathcal{P}(\llbracket n \rrbracket)$, a ranking $\widehat{\pi} \in \Gamma(A)$ drawn uniformly at random.

B. Experiments on the SUSHI dataset

The SUSHI dataset, described in [25] and available at http://www.kamishima.net, is composed of 5000 full rankings on a set of 10 sushi varieties. To generate a dataset of incomplete rankings of size 2 to 5, we first compute, for each full ranking, all its sub-rankings of size 5. Then each ranking of size 5 is censored into one of its sub-rankings of size $k \in \{2, \dots, 5\}$ drawn uniformly at random, k being equally drawn uniformly at random. We thus obtain a global dataset of $1,260,000$ incomplete rankings with size uniformly distributed on $\{2, \dots, 5\}$, for which we keep 80% as a training set and 20% as a test set. We evaluate the estimator $\tilde{p}_N$ but also its truncated versions to scales $k = 2$, 3, or 4, where the coefficients $\tilde{c}_\tau$ are put equal to 0 for $|\operatorname{supp}(\tau)| > k$, and compare them to the Plackett-Luce model (fitted via maximum likelihood estimation using the MM algorithm proposed in [5]) The results are shown on figure 4. They represent the empirical risks on the test set for the "plug-in" predictors of the five probabilistic models for the four loss functions. All plug-in predictors are computed exactly. As a baseline, the expectation and standard deviation of the uniformly random predictor are given in table I.

TABLE I
EXPECTATION AND STANDARD DEVIATION OF THE UNIFORMLY RANDOM CLASSIFIER FOR THE SUSHI DATASET

	$0-1$ loss	Kendall's tau	$l1$ distance	$l2$ distance
Expectation	0.8208	0.5000	0.6145	0.6156
Std. Dev.	6.6×10^{-4}	6.7×10^{-4}	7.1×10^{-4}	6.8×10^{-4}

In each case, the risk of worse model is lower than that of the uniformly random predictor by hundreds of standard deviations. This is surely explained by the fact that all statistical models manage to leverage information from historical data to make better predictions than random, and the amount of the difference is due to the large size of the test set $(252,000)$. Except for the Kendall's tau distance and the associated plug-in predictor, the multiresolution approach outperforms the Plackett-Luce model. An interesting observation is that for each loss function, the risk of the truncated multiresolution-based predictor decreases with the scale. This means that each scale contains a specific part of information that is useful to make better predictions. It shows in particular that reducing the observations to pairwise comparisons inherently degrades the available information, and proves the interest to exploit higher order information.

C. Experiments on the NETFLIX dataset

The NETFLIX dataset was issued for the Netflix Prize that took place between October 2, 2006 and September 21, 2009. The training set contains $100,480,507$ ratings given by $480,189$ users to $17,770$ movies. Each rating is a quadruplet $\langle$ `user, movie, date of grade, grade` $\rangle$, where the grade is an integer between 1 and 5. We use the training set to generate preference data, on the following simple paradigm: if a user gave respectively the grades g_a and g_b to movies a and b with $g_a > g_b$ then it means that she prefers movie a to movie b. More generally if she gave the grades $g_1 > \cdots > g_k$ to the movies $a_1, \dots, a_k$, her preference over the subset of movies $a_1, \dots, a_k$ is given by the ranking $a_1 \dots a_k$. As the grades are on a scale from 1 to 5, preferences take the form of incomplete rankings of maximum size 5. To generate the preference data, we consider for each user the list of the ratings she gave. As the average number of ratings per user is approximately 210, the brutal computation of all the possible preferences would require around $(210/5)^5 \times 500,000 = 92 \times 10^{12}$ operations, which is too costly. We therefore generate a sub-sample of this data with the following procedure.

1) For each user, we sort the list of the triplet $\langle$ `movie, date of grade, grade` $\rangle$ by chronological order.
2) We scroll the list and for each new movie, we draw a subset of m movies among the previously rated and compute all the possible preferences between them and the new movie.

This procedure mimics the way a user can rate movies: by comparing a new movie to some of the movies she previously

rated. As the m previously rated movies are drawn at random, some may have the same rate. In that case they generate two different preferences of smaller size. We choose $m = 4$ so that preferences of size 5 can appear. For each user, we keep the first 80% ratings for training and the 20% lasts for test. We then aggregate the data to obtain a training set and a test set of respectively $153,703,541$ and $38,665,610$ incomplete rankings.

We collected all the data in a distributed key/value framework and implemented the multiresolution-based estimator $\tilde{p}_N$ using map/reduce jobs, as described in section IV. For computation reasons, we only tested the predictive rule consisting in choosing the ranking with higher probability, for both the multiresolution-based estimator and the Plackett-Luce model. We nevertheless evaluated their performance through the four loss functions considered in the previous section. The results, as well as the expectation and standard deviation of the uniformly random classifier, are presented in table II.

TABLE II
RESULTS FOR THE NETFLIX DATASET

	$0-1$ loss	Kendall's tau	$l1$ distance	$l2$ distance
Expectation	0.7388	0.5000	0.6059	0.5867
Std. Dev.	6.5×10^{-5}	6.1×10^{-5}	6.6×10^{-5}	6.3×10^{-5}
Plackett-Luce	0.5579	0.3598	0.3934	0.3865
$\tilde{p}_N$	0.6042	0.3938	0.4425	0.4328

Again, both statistical models outperform by far the random classifier. Contrary to the SUSHI dataset, the Plackett-Luce model performs better than our estimator for all loss functions. This is surely due to the fact that through its rigid shape, the Plackett-Luce model captures more efficiently global effects on the full set of items, namely the average rank of a movie in any incomplete ranking. It is indeed highlighted in [26] that the tendencies of some movies to receive higher ratings than others capture much of the information in Netflix data. On the contrary the multiresolution-based estimator is best fitted to capture pure relative preferences effects. This experiment demonstrates nevertheless the good scalability of MRA of incomplete rankings and the overall good performance of its application to ranking prediction calls for further deepening and new applications.

VI. CONCLUSION

The analysis of preference data, as now collected and gathered through a variety of applications, raise computational challenges. Their representation and the modeling of their variability is far from straightforward, though essential when considering statistical tasks such as distribution estimation or predictive ranking. In this paper, we showed the relevance of MRA of incomplete rankings in this regard. The encouraging results we obtained also suggest several lines of further research, to investigate in particular how to combine such representations with preprocessing techniques (*e.g.* clustering methods) in order to design efficient collaborative filtering procedures, the distributions estimated from real data in practice being generally highly multimodal.

ACKNOWLEDGMENT

The authors would like to thank the NCF platform at Telecom SudParis for providing us with the computing infrastructure that permitted to carry out the experiments described in the present paper.

REFERENCES

[1] F. Ricci, Rokach, L., B. Shapira, and P. Kantor, *Recommender Systems Handbook*. Springer, 2011.

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

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

[4] R. L. Plackett, "The analysis of permutations," *Applied Statistics*, vol. 2, no. 24, pp. 193–202, 1975.

[5] D. R. Hunter, "MM algorithms for generalized Bradley-Terry models," *The Annals of Statistics*, vol. 32, pp. 384–406, 2004.

[6] J. Guiver and E. Snelson, "Bayesian inference for plackett-luce ranking models," in *ICML*, 2009.

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

[8] R. Kondor and M. S. Barbosa, "Ranking with kernels in Fourier space," in *COLT*, 2010, pp. 451–463.

[9] M. Sun, G. Lebanon, and P. Kidwell, "Estimating probabilities in recommendation systems," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 61, no. 3, pp. 471–492, 2012.

[10] S. Clémençon, J. Jakubowicz, and E. Sibony, "Multiresolution analysis of incomplete rankings," *ArXiv e-prints*, 2014.

[11] J. Bennett and S. Lanning, "The netflix prize," in *Proceedings of KDD Cup and Workshop 2007*, 2007.

[12] A. Ali and M. Meila, "Experiments with kemeny ranking: What works when?" *Mathematical Social Sciences*, vol. 64, no. 1, pp. 28 – 40, 2012.

[13] J.-Y. Audibert and A.Tsybakov, "Fast learning rates for plug-in classifiers," *Annals of statistics*, vol. 35, no. 2, pp. 608–633, 2007.

[14] S. Clémençon and S. Robbiano, "Minimax learning rates for bipartite ranking and plug-in rules," in *Proceedings of the International Conference in Machine Learning, ICML'11*, 2011.

[15] H. Azari Soufiani, W. Chen, D. C. Parkes, and L. Xia, "Generalized method-of-moments for rank aggregation," in *Advances in Neural Information Processing Systems 26*, 2013, pp. 2706–2714.

[16] R. Kondor and W. Dempsey, "Multiresolution analysis on the symmetric group," in *Neural Information Processing Systems 25*, 2012.

[17] R. Coifman and M. Maggioni, "Diffusion wavelets," *Applied and Computational Harmonic Analysis*, vol. 21, pp. 53–94, 2006.

[18] M. Gavish, B. Nadler, and R. R. Coifman, "Multiscale wavelets on trees, graphs and high dimensional data: theory and applications to semi supervised learning," in *International Conference on Machine Learning*, 2010, pp. 567–574.

[19] D. K. Hammond, P. Vandergheynst, and R. Gribonval, "Wavelets on graphs via spectral graph theory," *Applied and Computational Harmonic Analysis*, vol. 30, no. 2, pp. 129 – 150, 2011.

[20] I. Ram, M. Elad, and I. Cohen, "Generalized tree-based wavelet transform," *IEEE Transactions on Signal Processing*, vol. 59, no. 9, pp. 4199–4209, 2011.

[21] R. M. Rustamov, "Average interpolating wavelets on point clouds and graphs," *CoRR*, vol. abs/1110.2227, 2011.

[22] X. Jiang, L.-H. Lim, Y. Yao, and Y. Ye, "Statistical ranking and combinatorial Hodge theory," *Math. Program.*, vol. 127, no. 1, pp. 203–244, 2011.

[23] O. Dalal, S. H. Sengemedu, and S. Sanyal, "Multi-objective ranking of comments on web," in *Proceedings of the 21st international conference on World Wide Web*, ser. WWW '12, 2012, pp. 419–428.

[24] B. Osting, C. Brune, and S. Osher, "Enhanced statistical rankings via targeted data collection," in *Journal of Machine Learning Research, W&CP (ICML 2013)*, vol. 28 (1), 2013, pp. 489–497.

[25] T. Kamishima, "Nantonac collaborative filtering: recommendation based on order responses." in *KDD*. ACM, 2003, pp. 583–588.

[26] Y. Koren, "The bellkor solution to the netflix grand prize," 2009.