Statistical Inference for Incomplete Ranking Data:
The Case of Rank-Dependent Coarsening

Mohsen Ahmadi Fahandar [1]   **Eyke Hüllermeier** [1]   **Inés Couso** [2]

Abstract

We consider the problem of statistical inference for ranking data, specifically rank aggregation, under the assumption that samples are incomplete in the sense of not comprising all choice alternatives. In contrast to most existing methods, we explicitly model the process of turning a full ranking into an incomplete one, which we call the *coarsening process*. To this end, we propose the concept of *rank-dependent* coarsening, which assumes that incomplete rankings are produced by projecting a full ranking to a random subset of ranks. For a concrete instantiation of our model, in which full rankings are drawn from a Plackett-Luce distribution and observations take the form of pairwise preferences, we study the performance of various rank aggregation methods. In addition to predictive accuracy in the finite sample setting, we address the theoretical question of consistency, by which we mean the ability to recover a target ranking when the sample size goes to infinity, despite a potential bias in the observations caused by the (unknown) coarsening.

1. Introduction

The analysis of rank data has a long tradition in statistics, and corresponding methods have been used in various fields of application, such as psychology and the social sciences (Marden, 1995). More recently, applications in information retrieval and machine learning have caused a renewed interest in the analysis of rankings and topics such as "learning-to-rank" (Liu, 2011) and preference learning (Fürnkranz & Hüllermeier, 2011).

In most applications, the rankings observed are *incomplete* or *partial* in the sense of including only a subset of the underlying choice alternatives (subsequently referred to as

"items"), whereas no preferences are revealed about the remaining ones—pairwise comparisons can be seen as an important special case. Somewhat surprisingly, most methods for learning from ranking data, including methods for rank aggregation, simply ignore the process of turning a full ranking into an incomplete one. Or, they implicitly assume that the process is unimportant from a statistical point of view, because the subset of items observed is independent of the underlying ranking.

Obviously, this assumption is often not valid, as shown by practically relevant examples such as top-k observations. Motivated by examples of this kind, we propose the concept of *rank-dependent coarsening*, which assumes that incomplete rankings are produced by projecting a full ranking to a random subset of ranks. The notion of "coarsening" is meant to indicate that an incomplete ranking can be associated with a set of complete rankings, namely the set of its consistent extensions—set-valued data of that kind is also called "coarse data" in statistics (Heitjan & Rubin, 1991; Gill et al., 1997). The idea of coarsening is similar to the interpretation of partial rankings as "censored data" (Lebanon & Mao, 2008). The assumption of rank-dependent coarsening can be seen as orthogonal to standard marginalization, which acts on items instead of ranks (Rajkumar & Agarwal, 2014; Sibony et al., 2015).

In addition to introducing a general statistical framework for analyzing incomplete ranking data (Section 3), we outline several problems and learning tasks to be addressed in this framework. Learning of the entire model will normally not be feasible, even under restrictive assumptions on the coarsening. A specifically interesting question, therefore, is to what extent and in what sense successful learning is possible for methods that are agnostic of the coarsening process. We investigate this question, both practically (Section 6) and theoretically (Section 7), for several ranking methods (Section 5) and a concrete instantiation of our framework, in which full rankings are drawn from a Plackett-Luce distribution and observations take the form of pairwise preferences (Section 4). In particular, we are interested in the property of consistency, by which we mean the ability to recover a target ranking when the sample size goes to infinity, despite a potential bias in the observations caused by the coarsening.

---

[1]Paderborn University, Germany [2]University of Oviedo, Spain. Correspondence to: Eyke Hüllermeier <eyke@upb.de>.

2. Preliminaries and Notation

Let $\mathbb{S}_K$ denote the collection of rankings (permutations) over a set $U = \{a_1, \ldots, a_K\}$ of K items a_k, $k \in [K] = \{1, \ldots, K\}$. We denote by $\pi : [K] \longrightarrow [K]$ a complete ranking (a generic element of $\mathbb{S}_K$), where $\pi(k)$ denotes the position of the kth item a_k in the ranking, and by π^{-1} the ordering associated with a ranking, i.e., $\pi^{-1}(j)$ is the index of the item on position j. We write rankings in brackets and orderings in parentheses; for example, $\pi = [2, 4, 3, 1, 5]$ and $\pi^{-1} = (4, 1, 3, 2, 5)$ both denote the ranking $a_4 \succ a_1 \succ a_3 \succ a_2 \succ a_5$.

For a possibly incomplete ranking, which includes only some of the items, we use the symbol τ (instead of π). If the kth item does not occur in a ranking, then $\tau(k) = 0$ by definition; otherwise, $\tau(k)$ is the rank of the kth item. In the corresponding ordering, the missing items do simply not occur. For example, the ranking $a_4 \succ a_1 \succ a_2$ would be encoded as $\tau = [2, 3, 0, 1]$ and $\tau^{-1} = (4, 1, 2)$, respectively. We let $I(\tau) = \{k : \tau(k) > 0\} \subset [K]$ and denote the set of all rankings (complete or incomplete) by $\overline{\mathbb{S}}_K$.

An incomplete ranking τ can be associated with its set of linear extensions $E(\tau) \subset \mathbb{S}_K$, where $\pi \in E(\tau)$ if π is consistent with the order of items in $I(\tau)$, i.e., $(\tau(i) - \tau(j))(\pi(i) - \pi(j)) \geq 0$ for all $i, j \in I(\tau)$. An important special case is an incomplete ranking $\tau = \tau_{i,j} = (i, j)$ in the form of a pairwise comparison $a_i \succ a_j$ (i.e., $\tau(i) = 1$, $\tau(j) = 2$, $\tau(k) = 0$ otherwise), which is associated with the set of extensions

$$E(\tau) = E(a_i \succ a_j) = \{\pi \in \mathbb{S}_K : \pi(i) < \pi(j)\} \ .$$

Modeling an incomplete observation τ by the set of linear extensions $E(\tau)$ reflects the idea that τ has been produced from an underlying complete ranking π by some "coarsening" or "imprecisiation" process, which essentially consists of omitting some of the items from the ranking. $E(\tau)$ then corresponds to the set of all consistent extension π if nothing is known about the coarsening, except that it does not change the relative order of any items.

3. General Setting and Problems

The type of data we assume as observations is incomplete rankings $\tau \in \overline{\mathbb{S}}_K$. Statistical inference for this type of data requires a probabilistic model of the underlying data generating process, that is, a probability distribution on $\overline{\mathbb{S}}_K$.

3.1. A Stochastic Model for Incomplete Rankings

Recalling our idea of a coarsening process, it is natural to consider the data generating process as a two step procedure, in which a full ranking π is generated first and turned into an incomplete ranking τ afterward. We model this assumption in terms of a distribution on $\overline{\mathbb{S}}_K \times \mathbb{S}_K$, which assigns a degree of probability to each pair (τ, π). More specifically, we assume a parameterized distribution of the following form:

$$\boldsymbol{p}_{\theta,\lambda}(\tau, \pi) = \boldsymbol{p}_\theta(\pi) \cdot \boldsymbol{p}_\lambda(\tau \,|\, \pi) \tag{1}$$

Thus, while the generation of full rankings is determined by the distribution

$$\boldsymbol{p}_\theta : \mathbb{S}_K \longrightarrow [0, 1] \ , \tag{2}$$

the coarsening process is specified by a family of conditional probability distributions

$$\left\{\boldsymbol{p}_\lambda(\cdot \,|\, \pi) \,:\, \pi \in \mathbb{S}_K, \ \lambda \in \Lambda\right\} \ , \tag{3}$$

where λ collects all parameters of these distributions; $\boldsymbol{p}_{\theta,\lambda}(\tau, \pi)$ is the probability of producing the data $(\tau, \pi) \in \overline{\mathbb{S}}_K \times \mathbb{S}_K$. Note, however, that π is actually not observed.

3.1.1. RANK-DEPENDENT COARSENING

In its most general form, the coarsening process (3) is extremely rich, even if being restricted by the consistency assumption $\boldsymbol{p}_\lambda(\tau \,|\, \pi) = 0$ for $\pi \notin E(\tau)$. In fact, since the number of probabilities to be specified is of the order $2^K K!$, inference about λ will generally be difficult. Therefore, $\boldsymbol{p}_\lambda$ certainly needs to be restricted by further assumptions. Apart from practical reasons, such assumptions are also indispensable for successful learning. Otherwise, observations could be arbitrarily biased in favor or disfavor of items, so that an estimation of the underlying (full) preferences, as reflected by $\boldsymbol{p}_\theta$, becomes completely impossible. For example, the coarsening process may leave a ranking π unchanged whenever item a_1 is on the last position, and remove a_1 from π otherwise. Obviously, this item will then appear to have a very low preference.

As shown by this example, the estimation of preferences will generally be impossible unless the coarsening is somehow more "neutral". The assumption we make here is a property we call *rank-dependent* coarsening. A coarsening procedure is rank-dependent if the incompletion is only acting on *ranks* (positions) but not on *items*. That is, the procedure randomly selects a subset of ranks and removes the items on these ranks, independently of the items themselves. In other words, an incomplete observation τ is obtained by projecting a complete ranking π on a random subset of positions $A \in 2^{[K]}$, i.e., the family (3) of distributions $\boldsymbol{p}_\lambda(\cdot \,|\, \pi)$ is specified by a single measure on $2^{[K]}$. Or, stated more formally,

$$\boldsymbol{p}\big(\pi^{-1}(A) \,|\, \pi^{-1}\big) = \boldsymbol{p}\big(\sigma^{-1}(A) \,|\, \sigma^{-1}\big)$$

for all $\pi, \sigma \in \mathbb{S}^K$ and $A \subset [K]$, where $\pi^{-1}(A)$ denotes the projection of the ordering π^{-1} to the positions in A.

The assumption of rank-dependent coarsening can be seen as orthogonal to standard marginalization: while the latter projects a full ranking to a *subset of items*, the former

projects a ranking to a *subset of positions*. The practically relevant case of top-k observations is a special (degenerate) case of rank-dependent coarsening, in which

$$\boldsymbol{p}(A) = \begin{cases} 1 & \text{if } A = \{1, \ldots, k\} \\ 0 & \text{otherwise} \end{cases}$$

This model could be weakened in various ways. For example, instead of fixing the length of observed rankings to a constant, k could be a random variable. Or, one could assume that positions are discarded with increasing probability, though independently of each other; thus, the probability to observe a subset of items on ranks $A \subseteq \{1, \ldots, K\}$ is given by

$$P(A) = \prod_{i \in A} \lambda_i \cdot \prod_{j \notin A} (1 - \lambda_j).$$

The coarsening is then defined by the K parameters $\lambda_1 > \lambda_2 > \ldots > \lambda_K$.

3.2. Learning Tasks

Suppose a sample of (training) data $\mathcal{D} = \{\tau_1, \ldots, \tau_N\}$ to be given. As for statistical inference about the process (1), several problems could be tackled.

- The most obvious problem is to estimate the complete distribution on $\overline{\mathbb{S}}_K$, i.e., the parameters θ and λ. As already explained before, this will require specific assumptions about the coarsening.

- A somewhat weaker goal is to estimate the "precise part", i.e., the parameter θ. Indeed, in many cases, θ will be the relevant part of the model, as it specifies the preferences on items, whereas the coarsening is rather considered as a complication of the estimation. In this case, λ is of interest only in so far as it helps to estimate θ. Ideally, it would even be possible to estimate θ without any inference about λ, i.e., by simply ignoring the coarsening process.

- An even weaker goal is to estimate, not the parameter θ itself, but only an underlying "ground truth" ranking π^* associated with θ. Indeed, in the context of learning to rank, the ultimate goal is typically to predict a ranking, not necessarily a complete distribution. For example, the ranking π^* could be the mode of the distribution $\boldsymbol{p}_\theta$, or any other sort of representative statistics. This problem is especially relevant in practical applications such as rank aggregation, in which π^* would play the role of a consensus ranking.

Here, we are mainly interested in the third problem, i.e., the estimation of a ground-truth ranking π^*. Moreover, due to reasons of efficiency, we are aiming for an estimation technique that circumvents direct inference about λ, while

being robust in the sense of producing reasonably good results for a wide range of coarsening procedures.

4. Specific Setting and Problems

The development and analysis of methods is only possible for concrete instantiations of the setting introduced in the previous section. An instantiation of that kind will be proposed in this section. The first part of our data generating process, $\boldsymbol{p}_\theta$, will be modeled by the Plackett-Luce model (Plackett, 1975; Luce, 1959). To make the second part, $\boldsymbol{p}_\lambda$, manageable, we restrict observations to the practically relevant case of pairwise comparisons (i.e., incomplete rankings of length 2).

4.1. The Plackett-Luce Model

The Plackett-Luce (PL) model is parameterized by a vector $\theta = (\theta_1, \theta_2, \ldots, \theta_K) \in \Theta = \mathbb{R}_+^K$. Each θ_i can be interpreted as the weight or "strength" of the option a_i. The probability assigned by the PL model to a ranking represented by a permutation $\pi \in \mathbb{S}_K$ is given by

$$\boldsymbol{pl}_\theta(\pi) = \prod_{i=1}^{K} \frac{\theta_{\pi^{-1}(i)}}{\theta_{\pi^{-1}(i)} + \theta_{\pi^{-1}(i+1)} + \ldots + \theta_{\pi^{-1}(K)}} \quad (4)$$

Obviously, the PL model is invariant toward multiplication of θ with a constant $c > 0$, i.e., $\boldsymbol{pl}_\theta(\pi) = \boldsymbol{pl}_{c\theta}(\pi)$ for all $\pi \in \mathbb{S}_K$ and $c > 0$. Consequently, θ can be normalized without loss of generality (and the number of degrees of freedom is only $K - 1$ instead of K). Note that the most probable ranking, i.e., the mode of the PL distribution, is simply obtained by sorting the items in decreasing order of their weight:

$$\pi^* = \arg\max_{\pi \in \mathbb{S}_K} \boldsymbol{pl}_\theta(\pi) = \arg\operatorname{sort}_{k \in [K]} \{\theta_1, \ldots, \theta_K\}. \quad (5)$$

As a convenient property of PL, let us mention that it allows for an easy computation of marginals, because the marginal probability on a subset $U' = \{a_{i_1}, \ldots, a_{i_J}\} \subset U$ of $J \leq K$ items is again a PL model parametrized by $(\theta_{i_1}, \ldots, \theta_{i_J})$. Thus, for every $\tau \in \overline{\mathbb{S}}_K$ with $I(\tau) = U'$,

$$\boldsymbol{pl}_\theta(\tau) = \prod_{j=1}^{J} \frac{\theta_{\tau^{-1}(j)}}{\theta_{\tau^{-1}(j)} + \theta_{\tau^{-1}(j+1)} + \ldots + \theta_{\tau^{-1}(J)}}$$

In particular, this yields pairwise probabilities

$$p_{i,j} = \boldsymbol{pl}_\theta(\tau_{i,j}) = \frac{\theta_i}{\theta_i + \theta_j}, \quad (6)$$

where $\tau_{i,j} = (i, j)$ represents the preference $a_i \succ a_j$. This is the well-known Bradley-Terry-Luce model (Bradley & Terry, 1952), a model for the pairwise comparison of alternatives. Obviously, the larger θ_i in comparison to θ_j, the

higher the probability that a_i is chosen. The PL model can be seen as an extension of this principle to more than two items: the larger the parameter θ_i in (4) in comparison to the parameters θ_j, $j \neq i$, the higher the probability that a_i appears on a top rank.

4.2. Pairwise Preferences

If rank-dependent coarsening is restricted to the generation of pairwise comparisons, the entire distribution $\boldsymbol{p}_\lambda$ is specified by the set of $K(K-1)/2$ probabilities

$$\left\{ \lambda_{i,j} \,|\, 1 \leq i < j \leq K, \lambda_{i,j} \geq 0, \sum_{1 \leq i < j \leq K} \lambda_{i,j} = 1 \right\}, \quad (7)$$

where $\lambda_{i,j}$ denotes the probability that the ranks i and j are selected.

The problem of ranking based on pairwise comparisons has been studied quite extensively in the literature, albeit without taking coarsening into account, i.e., without asking where the pairwise comparisons are coming from (or implicitly assuming they are generated as marginals). Yet, worth mentioning is a recent study on *rank breaking* (Soufiani et al., 2014), that is, of the estimation bias (for models such as PL) caused by replacing full rankings in the training data by the set of all pairwise comparisons—our study of the bias caused by coarsening is very much in the same spirit.

4.3. The Data Generating Process

Combining the PL model (4) with the coarsening process (7), we obtain a distribution $\boldsymbol{q}$ on $\overline{\mathbb{S}}_K$ such that

$$\boldsymbol{q}(\tau_{i,j}) = q_{i,j} = \sum_{\pi \in E(a_i \succ a_j)} \boldsymbol{pl}_\theta(\pi) \, \lambda_{\pi(i),\pi(j)} \quad (8)$$

for pairwise preferences $\tau_{i,j} = (i,j)$, and $\boldsymbol{q}(\tau) = 0$ otherwise. Clearly, the pairwise probabilities (8) will normally not agree with the pairwise marginals (6). Instead, they may provide a biased view of the pairwise preferences between items. Please note, however, that the marginals $p_{i,j}$ are not directly comparable with the $q_{i,j}$, because the latter is a distribution on incomplete rankings ($\sum_{i,j} q_{i,j} = 1$) whereas the former is a set of marginal distributions ($p_{i,j} + p_{j,i} = 1$). Instead, $p_{i,j}$ should be compared to $q'_{i,j} = q_{i,j}/(q_{i,j} + q_{j,i})$, which is the probability that, in the coarsened model, a_i is observed as a winner, given it is paired with a_j.

As an illustration, consider a concrete example with $K = 3$, $\theta = (14, 5, 1)$, and degenerate coarsening distribution specified by $\lambda_{1,2} = 1$ (top-2 selection). One easily derives the probabilities of pairwise marginals (6) and coarsened (top-2) observations (8) as follows:

i, j	$1, 2$	$1, 3$	$2, 3$	$2, 1$	$3, 1$	$3, 2$
$p_{i,j}$	$\frac{840}{1140}$	$\frac{1064}{1140}$	$\frac{950}{1140}$	$\frac{300}{1140}$	$\frac{76}{1140}$	$\frac{190}{1140}$
$q_{i,j}$	$\frac{665}{1140}$	$\frac{133}{1140}$	$\frac{19}{1140}$	$\frac{266}{1140}$	$\frac{42}{1140}$	$\frac{15}{1140}$
$q'_{i,j}$	$\frac{665}{931}$	$\frac{133}{175}$	$\frac{19}{34}$	$\frac{266}{931}$	$\frac{42}{175}$	$\frac{15}{34}$

While the $p_{i,j}$ are completely coherent with a PL model (namely $\boldsymbol{pl}_\theta$ with $\theta = (14, 5, 1)$), the $q_{i,j}$ and $q'_{i,j}$ no longer are.

A special case where coarsening is guaranteed to not introduce any bias is the uniform distribution $\lambda \equiv 2/(K^2 - K)$. In this case, random projection to ranks effectively coincides with random selection to items.

4.4. Problems

Of course, in spite of the inconsistency of the pairwise observations in the above example, a PL model could still be estimated, for example using the maximum likelihood principle. The corresponding estimate $\hat{\theta}$ will also yield an estimate $\hat{\pi}$ of the target ranking π^* (which is simply obtained by sorting the items a_i in decreasing order of the estimated scores θ_i). As already said, there is little hope that $\hat{\theta}$ could be an unbiased estimate of θ; instead $\hat{\theta}$ will necessarily be biased. There is hope, however, to recover the target ranking π^*. Indeed, the ranking will be predicted correctly provided $\hat{\theta}$ is comonotonic with θ, i.e., $(\hat{\theta}_i - \hat{\theta}_j)(\theta_i - \theta_j) > 0$ for all $i, j \in [K]$. Roughly speaking, a small bias in the estimate can be tolerated, as long as the order of the parameters is preserved.

Obviously, these considerations are not restricted to the PL model. Instead, any method for aggregating pairwise comparisons into an overall ranking can be used to predict π^*. This leads us to the following questions:

- Practical performance: What is the performance of a rank aggregation method in the finite sample setting, i.e., how close is the prediction $\hat{\pi}$ to the ground truth π^*? How is the performance influenced by the coarsening process?

- Consistency: Is a method consistent in the sense that π^* is recovered (with high probability) with an increasing sample size $N \to \infty$, either under specific assumptions on the coarsening process, or perhaps even regardless of the coarsening (i.e., only assuming the property of rank-dependence)?

5. Rank Aggregation Methods

In this section, we discuss different rank aggregation methods that operate on pairwise data, categorized according to the some basic principles. To this end, let us define the comparison matrix C with entries $c_{i,j}$, where $c_{i,j}$ denotes the

number of wins of a_i over a_j, i.e., the number of times the preference $a_i \succ a_j$ is observed. Correspondingly, we define the probability matrix $\hat{P}$ with entries $\hat{p}_{i,j} = \frac{c_{i,j}}{c_{i,j}+c_{j,i}}$, which can be seen as estimates of the winning probabilities $p_{i,j}$ (often also interpreted as weighted preferences). All rank aggregation methods produce rankings $\hat{\pi}$ based on either matrix C or $\hat{P}$.

5.1. Statistical Estimation

5.1.1. BRADLEY-TERRY-LUCE MODEL (BTL)

The Bradley-Terry-Luce model is well-known in the literature on discrete choice (Bradley & Terry, 1952). It starts from the parametric model (6) of pairwise comparisons, i.e., the marginals of the PL model, and estimates the parameters by likelihood maximization:

$$\hat{\theta} \in \arg\max_{\theta \in \mathbb{R}^K} \prod_{1 \leq i \neq j \leq K} \left(\frac{\theta_i}{\theta_i + \theta_j} \right)^{c_{i,j}}$$

The predicted ranking is obtained by sorting items according to their estimated strengths: $\hat{\pi} = \arg\text{sort}(\hat{\theta})$.

As already explained, coarsening of rankings may cause a bias in the number of observed pairwise preferences. In particular, it may happen that some (pairs of) items are observed much more often than others. Therefore, in addition to the BTL problem as formalized above, we also consider the same problem with relative winning frequencies $\hat{p}_{i,j}$ instead of absolute frequencies $c_{i,j}$; we call this approach BTL(R).

5.1.2. LEAST SQUARES/HODGERANK (LS)

The HodgeRank algorithm (Jiang et al., 2011) is based on a least squares approach. First, the probability matrix $\hat{P}$ is mapped to a matrix X as follows:

$$X_{i,j} = \begin{cases} \log\left(\frac{\hat{p}_{i,j}}{\hat{p}_{j,i}} \right) & \text{if } i \neq j \text{ and } \hat{p}_{j,i} \in (0,1), \\ 0 & \text{otherwise} \end{cases}$$

Then, $\hat{\pi} = \arg\text{sort}(\theta^*)$, where

$$\theta^* \in \arg\min_{\theta \in \mathbb{R}^K} \sum_{(i,j) \in \mathcal{E}} \left((\theta_j - \theta_i) - X_{i,j} \right)^2,$$

and $\mathcal{E} = \left\{ (i,j) \mid 1 \leq i < j \leq K, \ X_{i,j} \neq 0 \right\}$.

5.2. Voting Methods

5.2.1. BORDA COUNT (BORDA)

Borda (Borda, 1781) is a scoring rule that sorts items according to the sum of weighted preferences or "votes" in favor of each item:

$$\hat{\pi} = \arg\text{sort}\{s_1, \ldots, s_K\} \ ,$$

where $s_i = \sum_{i=1}^{K} \hat{p}_{i,j}$.

5.2.2. COPELAND (CP)

Copeland (Copeland, 1951) works in the same way as Borda, except that scores are derived from binary instead of weighted votes:

$$s_i = \sum_{i=1}^{K} \mathbb{I}\left(\hat{p}_{i,j} > \frac{1}{2} \right) \ .$$

5.3. Spectral Methods

The idea of deriving a consensus ranking from the stationary distribution of a suitably constructed Markov chain has been thoroughly studied in the literature (Seeley, 1949; Vigna, 2009; Brin & Page, 1998). The corresponding Markov chain with transition probabilities Q is defined by the pairwise preferences. Then, if Q is an irreducible, aperiodic Markov chain, the stationary distribution $\bar{\pi}$ can be computed, and the predicted ranking is given by $\hat{\pi} = \arg\text{sort}(\bar{\pi})$.

5.3.1. RANK CENTRALITY (RC)

The rank centrality algorithm (Negahban et al., 2012) is based on the following transition probabilities:

$$Q_{i,j} = \begin{cases} \dfrac{1}{K}\, \hat{p}_{i,j} & \text{if } i \neq j \\ 1 - \dfrac{1}{K} \displaystyle\sum_{k \neq i} \hat{p}_{k,i} & \text{if } i = j \end{cases}$$

5.3.2. MC2 AND MC3

Dwork et al. (2001) introduce four spectral ranking algorithms, two of which we consider for our study (namely MC2 and MC3), translated to the setting of pairwise preferences. For MC2, the transition probabilities are given as follows:

$$Q_{i,j} = \begin{cases} \dfrac{1}{\sum_{j=1}^{K} \hat{p}_{i,j}}\, \hat{p}_{j,i} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

The MC3 algorithm is based on the transition probabilities

$$Q_{i,j} = \begin{cases} \dfrac{1}{deg(a_i)}\, \hat{p}_{i,j} & \text{if } i \neq j \\ 1 - \dfrac{1}{deg(a_i)} \displaystyle\sum_{k \neq i} \hat{p}_{k,i} & \text{if } i = j \end{cases}$$

where

$$deg(a_i) = \max\left(\sum_{i=1}^{K} \mathbb{I}(\hat{p}_{i,j} > 0), \sum_{i=1}^{K} \mathbb{I}(\hat{p}_{j,i} > 0) \right).$$

5.4. Graph-based Methods

5.4.1. FEEDBACK ARC SET (FAS)

If the $c_{i,j}$ are interpreted as (weighted) preferences, the degree of inconsistency of ranking a_i before a_j is naturally quantified in terms of $c_{j,i}$. Starting from a formalization in terms of a graph whose nodes correspond to the items and whose edges are labeled with the weighted preferences, the weighted feedback arc set problem (Saab, 2001; Fomin et al., 2010) is to find the ranking that causes the lowest sum of penalties:

$$\hat{\pi} = \arg\min_{\pi \in \mathbb{S}_K} \sum_{(i,j): \pi(i) < \pi(j)} c_{j,i}$$

For the same reason as in the case of BTL, we also consider the FAS problem with edge weights given by relative winning frequencies $\hat{p}_{i,j}$ and binary preferences $\mathbb{I}(\hat{p}_{i,j} > 1/2)$ instead of absolute frequencies $c_{i,j}$; we call the former approach FAS(R) and the latter FAS(B).

5.5. Pairwise Coupling

A common approach to multi-class classification is the all-pairs decomposition, in which one binary classifier $h_{i,j}$ is trained for each pair of classes a_i and a_j (Fürnkranz, 2002). At prediction time, each classifier produces a prediction, which can be interpreted as a vote, or weighted vote $\hat{p}_{i,j}$ in case of a probabilistic classifier, in favor of item a_i over a_j. The problem of combining these predictions into an overall prediction for the multi-class problem is also called *pairwise coupling* (Hastie & Tibshirani, 1998).

To the best of our knowledge, pairwise coupling has not been used for rank aggregation so far. In fact, the original purpose of this technique is not to rank items but merely to identify a single winner. Nevertheless, since coupling methods are eventually based on scoring items, they can be generalized for the purpose of ranking in a straightforward way. Indeed, they have been used for that purpose in the context of label ranking (Hüllermeier et al., 2008).

5.5.1. METHOD BY HASTIE AND TIBSHIRANI (HT)

Hastie & Tibshirani (1998) tackle the problem in the following way: Given relative frequencies $\hat{P}$, they suggest to find the probability vector $\boldsymbol{p} = (p_1, \dots, p_K)$ that minimizes the (weighted) Kullback-Leibler (KL) distance

$$\ell(\boldsymbol{p}) = \sum_{i<j} n_{i,j} \left[\hat{p}_{i,j} \log \frac{\hat{p}_{i,j}}{\mu_{i,j}} + (1 - \hat{p}_{i,j}) \log \frac{1 - \hat{p}_{i,j}}{1 - \mu_{i,j}} \right]$$

between $\hat{p}_{i,j}$ and $\mu_{i,j} = \frac{p_i}{p_i + p_j}$, where $n_{i,j} = c_{i,j} + c_{j,i}$. To this end, the problem is formulated as a fixed point problem and solved using an iterative algorithm. Once $\boldsymbol{p}$ is obtained, the predicted ranking is determined by $\hat{\pi} = \arg\mathrm{sort}(\boldsymbol{p})$.

5.5.2. METHOD BY PRICE ET AL. (PRICE)

Price et al. (1994) propose the following parameter estimation for each $i \in [K]$:

$$\hat{\theta}_i = \frac{1}{\left(\sum_{j \neq i} \frac{1}{\hat{p}_{i,j}} \right) - (K-2)}$$

Then, the predicted ranking is given by $\hat{\pi} = \arg\mathrm{sort}(\hat{\theta})$.

5.5.3. METHOD BY TING-FAN WU ET AL. (WU1, WU2)

Wu et al. (2004) propose two methods. The first one (WU1) is based on the Markov chain approach with transition matrix

$$Q_{i,j} = \begin{cases} \hat{p}_{i,j}/(K-1) & \text{if } i \neq j \\ \sum_{s \neq i} \hat{p}_{i,s}/(K-1) & \text{if } i = j \end{cases} .$$

Once the stationary distribution $\bar{\pi}$ of Q is obtained, the predicted ranking is given by $\hat{\pi} = \arg\mathrm{sort}(\bar{\pi})$. In their second approach (WU2), the following optimization problem is proposed:

$$\min_{\theta} 2\theta^T Q \theta,$$

where

$$Q_{i,j} = \begin{cases} -\hat{p}_{i,j}\hat{p}_{j,i} & \text{if } i \neq j \\ \sum_{s \neq i} \hat{p}_{s,i}^2 & \text{if } i = j \end{cases} .$$

Once the optimization is solved, the predicted ranking is $\hat{\pi} = \arg\mathrm{sort}(\theta)$.

6. Practical Performance

6.1. Synthetic Data

To investigate the practical performance of the methods presented in the previous section, we first conducted controlled experiments with synthetic data, for which the ground truth π^* is known. To this end, data in the form of pairwise observations was generated according to (1) with different distributions $\boldsymbol{p}_\theta$ and $\boldsymbol{p}_\lambda$, predictions $\hat{\pi}$ were produced and compared with π^* in terms of the Kendall distance, i.e., the number of pairwise inversions:

$$\sum_{1 \leq i < j \leq K} \mathbb{I}\left[\mathrm{sign}(\pi^*(i) - \pi^*(j)) \neq \mathrm{sign}(\hat{\pi}(i) - \hat{\pi}(j)) \right].$$

For each setting, specified by parameters K, θ, λ, we are interested in the expected performance of a method as a function of the sample size N, which was approximated by averaging over 500 simulation runs.

6.1.1. PL DISTRIBUTION

In a first series of experiments, synthetic data was produced for $K \in \{3, 4, 5, 7\}$, $\boldsymbol{p}_\theta$ the PL model with ground truth

parameter $\theta \in \mathbb{R}_+^K$, and coarsening rankings by projecting to all possible pairs of ranks (i.e., using a degenerate distribution $\boldsymbol{p}_\lambda$ with $\lambda_{i,j} = 1$ for some $1 \leq i < j \leq K$). As a baseline, we also produced the performance of each method for the case of *full* pairwise information about a ranking, i.e., adding all pairwise preferences to the data (instead of only $a_{\pi^{-1}(i)} \succ a_{\pi^{-1}(j)}$) that can be extracted from a ranking π.

Due to space restrictions, the results are only shown in the supplementary material. The following observations can be made:

- Comparing the results for learning from incomplete data with the baseline, it is obvious that coarsening comes with a loss of information and makes learning more difficult.

- The difficulty of the learning task increases with decreasing distance $|i - j|$ between observed ranks and is most challenging for the case of observing neighboring ranks $(i, i+1)$.

- All methods perform rather well, although FAS is visibly worse while BTL is a bit better than the others. On the one side, this could be explained by the fact that the BTL model is consistent with the PL model, as it corresponds to the marginals of $\boldsymbol{p}_\theta$. On the other side, like all other methods, BTL is agnostic of the coarsening $\boldsymbol{p}_\lambda$; from this point of view, the strong performance is indeed a bit surprising. As a side remark, BTL(R) does not improve on BTL.

- While FAS is on a par with FAS(R), FAS(B) tends to do slightly better, especially for a larger number of items. However, as already said, FAS performs generally worse than the others.

In another set of experiments, we examined the *averaged* performance of methods over all coarsening positions. The results are again shown in the supplementary material. It is noticeable that, as the number of items increases, the performance of the FAS-based approaches decreases. Moreover, as pointed out earlier, the BTL performs moderately better than other approaches.

6.1.2. MALLOWS DISTRIBUTION

In a second series of experiments, we replaced the PL distribution with another well-known distribution on rankings, namely the Mallows distribution. Thus, data is now generated with $\boldsymbol{p}_\theta$ in (1) given by Mallows instead of PL. This experiment serves as a kind of sensitivity analysis, especially for those methods that (explicitly or implicitly) rely on the assumption of PL.

The Mallows model (Mallows, 1957) is parametrized by a reference ranking π^* (which is the mode) and a dispersion
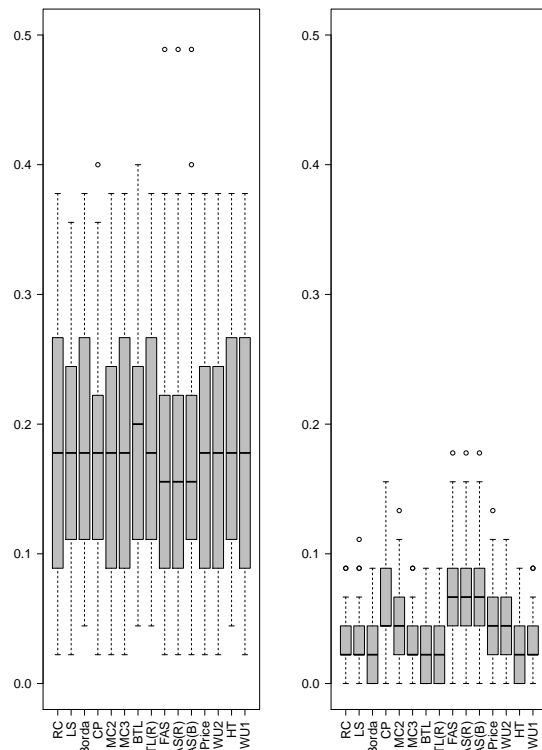

Figure 1. Results for the sushi data in terms of normalized Kendall distance (one boxblot per method), experiment (a) on the left and (b) on the right.

parameter ϕ, i.e., $\theta = (\pi^*, \phi)$:

$$\boldsymbol{p}_{\pi^*,\phi}(\pi) = \frac{1}{Z(\phi)} \exp\big(-\phi D(\pi, \pi^*)\big),$$

where $D(\pi, \pi^*)$ is the Kendall distance and $Z(\phi)$ a normalization constant.

The results and observations for this series of experiments are quite similar to those for the PL model. What is notable, however, is a visible drop in performance for BTL, whereas Copeland ranking now performs much better than all other algorithms. Furthermore, FAS(B) is significantly better than FAS and FAS(R). These results might be explained by the ordinal nature of the Mallows model, which is arguably better fit by methods based on binary comparisons than by score-based approaches.

6.2. Real Data

To compare the methods on a real world data set, we used the Sushi data (Kamishima, 2003) that contains the preferences (full rankings) of 5000 people over 10 types of sushi.

For this data set, there is no obvious ground truth π^*. There-

fore, we define a separate target for each data set individually, which is the ranking produced by that method on the full set of pairwise comparisons that can be extracted from the training data. The distance between this ranking and the one predicted for coarsened data can essentially be seen as a measure of the loss of information caused by coarsening. We conduct this experiment for (a) degenerate coarsening where $\lambda_{i,j} = 1$ for some (i, j) and (b) random coarsening where $\lambda_{i,j} = 2/(K^2 - K)$ for all (i, j). Note that, while the information content is the same in both cases (one pairwise comparison per customer), random coarsening does not introduce any bias, as opposed to degenerate coarsening.

The results (distributions for 100 repetitions) are shown in Figure 1. Again, all methods are more or less on a par. However, as expected, random coarsening does indeed lead to better estimates, again suggesting that "real" coarsening indeed makes learning harder.

7. Consistency

Recall the (specific) setting we introduced in Section 4, in which full rankings are generated according to a PL model with parameter $\theta = (\theta_1, \ldots, \theta_K)$, and the coarsening corresponds to a projection to a random pair of ranks; for technical reasons, we subsequently assume $\theta_i \neq \theta_j$ for $i \neq j$. Also recall that we denote by $p_{i,j}$ the probability of a ranking π in which a_i precedes a_j, and by $q_{i,j}$ the probability to observe the preference $a_i \succ a_j$ after coarsening. According to PL, we have $p_{i,j} = \theta_i/(\theta_i + \theta_j)$, and the ground truth ranking π^* is such that

$$(\pi(i) < \pi(j)) \Leftrightarrow (\theta_i > \theta_j) \Leftrightarrow (p_{i,j} > 1/2)$$

Finally, we denote by $\hat{p}_{i,j}$ the estimate of the pairwise preference (probability) $p_{i,j}$. If not stated differently, we always assume estimates to be given by relative frequencies, i.e., $\hat{p}_{i,j} = w_{i,j}/(w_{i,j} + w_{j,i})$, with $w_{i,j}$ the observed number of preferences $a_i \succ a_j$.

Definition 1: Let $\hat{\pi}_N$ denote the ranking produced as a prediction by a ranking method on the basis of N observed (pairwise) preferences. The method is consistent if $p(\hat{\pi}_N = \pi^*) \to 1$ for $N \to \infty$.

The proofs of the following results are given in the supplementary material.

Lemma 2: Let us consider a probability measure p_θ over $\mathbb{S}_K$. Consider $q_{i,j} = \sum_{\pi \in E(a_i \succ a_j)} p_\theta(\pi)\lambda_{\pi(i),\pi(j)}, \forall i \neq j$. (The model (8) with $p_\theta(\pi)$ not necessarily PL). If $p_\theta(\pi) \geq p_\theta(\pi_{i,j})$ for all $\pi \in E(a_i \succ a_j)$, then $q_{i,j} > q_{j,i}$.

Lemma 3: Assume the model (8), $\theta_i \neq \theta_j$ for $i \neq j$, and $\theta_i > 0$ for all $i \in [K]$. The coarsening (7) is order-preserving for PL in the sense that $p_{i,j} > 1/2$ if and only if $q'_{i,j} > 1/2$, where $q'_{i,j} = q_{i,j}/(q_{i,j} + q_{j,i})$.

The last result is indeed remarkable: Although coarsening will bias the pairwise probabilities $p_{i,j}$, the "binary" preferences will be preserved in the sense that $\text{sign}(p_{i,j} - 1/2) = \text{sign}(q'_{i,j} - 1/2)$. Indeed, the result heavily exploits properties of the PL distribution and does not hold in general. For example, consider a distribution p on $\mathbb{S}_3$ such that $p([1, 2, 3]) = 0.8$, $p([3, 1, 2]) = p([3, 2, 1]) = 0.1$. Then, with a coarsening (7) such that $\lambda_{2,3} = 1$, we have $p_{1,2} = p_{1,3} = 0.8$, but $q'_{1,2} = q'_{1,3} = 0$.

Lemma 4: Assume the model (8), $\theta_i \neq \theta_j$ for $i \neq j$, and $\theta_i > 0$ for all $i \in [K]$. Let us take an arbitrarily small $\epsilon^* > 0$. There exists $N_0 \in \mathbb{N}$ such that $\theta_i > \theta_j$ if and only if $\hat{p}_{i,j} > 1/2$ for all $i, j \in [K]$, with probability at least $1 - \epsilon^*$, after having observed at least N_0 preferences.

Theorem 5: Copeland ranking is consistent.

Theorem 6: FAS, FAS(R), and FAS(B) are consistent.

Our experimental results so far suggest that consistency does not only hold for Copeland and FAS, but also for most other methods (including BTL), and hence that rank-dependent coarsening is indeed somehow "good-natured". Anyway, for these cases, the proofs are still pending.

8. Summary and Conclusion

In this paper, we addressed the problem of learning from incomplete ranking data and advocated an explicit consideration of the process of turning a full ranking into an incomplete one—a step that we referred to as "coarsening". To this end, we proposed a suitable probabilistic model and introduced the property of rank-dependent coarsening, which can be seen as orthogonal to standard marginalization: while the latter projects a ranking to a subset of items, the former projects to a subset of ranks.

First experimental and theoretical results suggest that agnostic learning can be successful under rank-dependent coarsening: even if ignorance of the coarsening may lead to biased parameter estimates, the ranking task itself can still be solved properly. This applies at least to the specific setting that we considered, namely rank aggregation based on pairwise preferences, with Plackett-Luce (or Mallows) as an underlying distribution.

Needless to say, this paper is only a first step. Many questions are still open, for example regarding the consistency of ranking methods, not only for the specific setting considered here but even more so for generalizations thereof. In addition to theoretical problems of that kind, we are also interested in practical applications such as "crowdordering" (Matsui et al., 2014; Chen et al., 2013), in which coarsening could play an important role.

References

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

Bradley, R.A. and Terry, M.E. The rank analysis of incomplete block designs I. The method of paired comparisons. *Biometrika*, 39:324–345, 1952.

Brin, S. and Page, L. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the 7th International Conference on World Wide Web*, pp. 107–117, Amsterdam, The Netherlands, 1998. Elsevier Science Publishers B. V.

Chen, X. Bennett, P.N. Collins-Thompson, K. and Horvitz, E. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining*, pp. 193–202, Rome, Italy, 2013.

Copeland, A. H. A 'reasonable' social welfare function. In *Seminar on Mathematics in Social Sciences, University of Michigan*, 1951.

Dwork, C., Kumar, R., Naor, M., and Sivakumar, D. Rank aggregation methods for the web. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pp. 613–622, New York, USA, 2001. ACM.

Fomin, F.V., Lokshtanov, D., Raman, V., and Saurabh, S. Fast local search algorithm for weighted feedback arc set in tournaments. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, AAAI'10, pp. 65–70. AAAI Press, 2010.

Fürnkranz, J. Round robin classification. *Journal of Machine Learning Research*, 2:721–747, 2002.

Fürnkranz, J. and Hüllermeier, E. *Preference Learning*. Springer-Verlag, 2011.

Gill, R. D., Laan, M. J., and Robins, J. M. Coarsening at random: Characterizations, conjectures, counter-examples. In *Proceedings of the 1st Seattle Symposium in Biostatistics: Survival Analysis*, pp. 255–294. Springer US, New York, 1997.

Hastie, T. and Tibshirani, R. Classification by pairwise coupling. In *Advances in Neural Information Processing Systems (NIPS'98)*. The MIT Press, 1998.

Heitjan, D.F. and Rubin, D.B. Ignorability and coarse data. *The Annals of Statistics*, 19(4):2244–2253, 1991.

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

Jiang, X., Lim, L., Yao, Y., and Ye, Y. Statistical ranking and combinatorial hodge theory. *Mathematical Programming*, 127(1):203–244, 2011.

Kamishima, T. Nantonac collaborative filtering: Recommendation based on order responses. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pp. 583–588, New York, USA, 2003. ACM.

Lebanon, G. and Mao, Y. Nonparametric modeling of partially ranked data. *Journal of Machine Learning Research*, 9:2401–2429, 2008.

Liu, T.Y. *Learning to Rank for Information Retrieval*. Springer-Verlag, 2011.

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

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

Marden, J.I. *Analyzing and Modeling Rank Data*. London, New York, 1995.

Matsui, T. Baba, Y. Kamishima, T. and Kashima, H. Crowdordering. In *Proceedings of the 18th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pp. 336–347, Tainan, Taiwan, 2014.

Negahban, S., Oh, S., and Shah, D. Iterative ranking from pair-wise comparisons. In *Advances in Neural Information Processing Systems (NIPS'12)*, pp. 2474–2482, 2012.

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

Price, D., Knerr, S., Personnaz, L., and Dreyfus, G. Pairwise neural network classifiers with probabilistic outputs. In *Proceedings of the 7th International Conference on Neural Information Processing Systems (NIPS'94)*, pp. 1109–1116, Cambridge, MA, USA, 1994. MIT Press.

Rajkumar, A. and Agarwal, S. A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In *Proceedings of the 31th International Conference on Machine Learning*, pp. 118–126, Beijing, China, 2014.

Saab, Y. A fast and effective algorithm for the feedback arc set problem. *Journal of Heuristics*, 7(3):235–250, May 2001.

Seeley, J. R. The net of reciprocal influence; a problem in treating sociometric data. *Canadian Journal of Psychology*, 3(4):234–240, Dec 1949.

Sibony, E. Clémençon, S. and Jakubowicz, J. MRA-based statistical learning from incomplete rankings. In *Proceedings of the 32th International Conference on Machine Learning*, pp. 1432–1441, Lille, France, 2015.

Soufiani, H.A. Parkes, D.C. and Xia, L. Computing parametric ranking models via rank-breaking. In *Proceedings of the 31th International Conference on Machine Learning*, Beijing, China, 2014.

Vigna, S. Spectral ranking. *CoRR*, abs/0912.0238, 2009.

Wu, T.F., Lin, C.J., and Weng, R.C. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5:975–1005, December 2004.