Journal of Machine Learning Research 9 (2008) 2401-2429 Submitted 7/07; Revised 5/08; Published 10/08

Non-Parametric Modeling of Partially Ranked Data

Guy Lebanon LEBANON@CC.GATECH.EDU
College of Computing
Georgia Institute of Technology
Atlanta, GA

Yi Mao YMAO@ECE.PURDUE.EDU
School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN

Editor: Tommi Jaakkola

Abstract

Statistical models on full and partial rankings of n items are often of limited practical use for large n due to computational consideration. We explore the use of non-parametric models for partially ranked data and derive computationally efficient procedures for their use for large n. The derivations are largely possible through combinatorial and algebraic manipulations based on the lattice of partial rankings. A bias-variance analysis and an experimental study demonstrate the applicability of the proposed method.

Keywords: ranked data, partially ordered sets, kernel smoothing

1. Introduction

Rankers such as people, search engines, and classifiers, output full or partial rankings representing preference relations over n items or alternatives. For example in the case of $m = 6$ rankers issuing full or partial preferences over $n = 3$ items a possible data set is

$$3 \prec 1 \prec 2, \quad 3 \prec 2 \prec 1, \quad 1 \prec 3 \prec 2, \quad 1 \prec \{2,3\}, \quad 3 \prec \{1,2\}, \quad \{2,3\} \prec 1. \tag{1}$$

The first three expressions in (1) correspond to full rankings while the last three expressions correspond to partial rankings (the numbers correspond to items and the $\prec$ symbol corresponds to a preference relation). While it is likely that some rankings will contradict others, it is natural to assume that the data in (1) was sampled iid from some distribution p over rankings. The goal of this paper is to study non-parametric methods for the estimation of p based on data sets such as (1) in the case of large n.

Often, ranked data is not inherently associated with numeric score information. In other cases, numeric scores are available but are un-calibrated and cannot be compared to each other. For example, the assignment of numeric scores by people to items or alternatives is un-calibrated as each person has his or her own notion of what constitutes a certain numeric score. On the other hand, a preference of one item or alternative over another reflects a binary choice that is directly comparable across rankers. Thus, even in cases where numeric scores exist, modeling the scoreless preferences may achieve higher modeling accuracy.

Despite this motivating observation, modeling ranked data is less popular than modeling the existing numeric scores, or even made-up numeric scores in case the true scores are unavailable (such is the case with the frequently used Borda count). The main reason for this is that rankings over a large number of items n reside in an extremely large discrete space whose modeling often requires intractable computation.

Previous attempts at modeling ranked data have been mostly parametric and often designed to work with fully ranked data (Marden, 1996). Non-parametric modeling of fully ranked data has been recently addressed in the context of multi-object tracking (Kondor et al., 2007; Huang et al., 2008). They focus on maintaining and updating a distribution over permutations by a low frequency approximation of the distribution. Such an approximation results from a spectral decomposition of functions on the symmetric group on n items (Diaconis, 1988) and is essential for efficient probabilistic inference.

Most of aforementioned approaches are unsuitable for modeling partial rankings for medium and large n due to the computational difficulties of handling a probability space of size $n!$. The few possible exceptions (Critchlow, 1985; Marden, 1996) are usually more ad-hoc and do not correspond to an underlying permutation model making them ill suited to handle partial rankings of different types. In fact, most of the ranked data analyzed in the literature are limited to $n \leq 15$ and usually even $n \leq 5$ such as the popular APA election data set.

On the other hand, there has been a recent increase in data sets containing partial or full rankings for large n. Examples include (i) web-search data such as TREC[1] where n may be thought of as corresponding to the number of web-pages or approaching $+\infty$, (ii) movie review data sets such as the Netflix data set[2] where $n \approx 18000$ and MovieLens[3] where $n = 1682$, and (iii) multi-label text document data sets such as OHSUMED[4] where $n = 4904$ and Reuters RCV1[5] where $n = 103$. More details on how these data sets correspond to partial rankings may be found in Section 2.

These data sets and others lead to a growing number of somewhat ad-hoc but computationally efficient rank aggregation techniques. The techniques, developed primarily within the computer science community, are often non-probabilistic and output a single ranking summarizing the data. Unfortunately, such a summary ranking, while being useful, does not provide the data analysis capabilities offered by a full probabilistic model.

The main contribution of this paper is in proposing and studying a non-parametric estimator based on kernel smoothing for the estimation of the population distribution p. Some properties of the estimator are listed below. We are not aware of any other non-trivial estimator of p that satisfies these requirements, in particular for the case of large n.

(1) Estimate p based on full as well as partial rankings.

(2) The resulting estimate $\hat{p}$ should assign probabilities to full and partial rankings in a coherent and contradiction-free manner (described in Section 4).

(3) Estimate p based on partial rankings of different types (defined in Section 2).

1. TREC can be found at `http://trec.nist.gov/`.
2. Netflix can be found at `http://www.netflixprize.com/`.
3. MovieLens can be found at `http://www.grouplens.org/node/12/`.
4. OHSUMED can be found at `http://trec.nist.gov/data/t9_filtering/`.
5. RCV1 can be found at `http://trec.nist.gov/data/reuters/reuters.html/`.

(4) Statistical consistency $\hat{p} \xrightarrow{p} p$ as both the number of samples m and the number of items n grow to infinity.

(5) Statistical accuracy of $\hat{p}$ can be slow for fully ranked data but should be accelerated when restricted to simpler partial rankings.

(6) Obtaining the estimate $\hat{p}$ and using it to compute probabilities $\hat{p}(A)$ of partial rankings should be computationally feasible, even for large n.

All 6 properties above are crucial in the large n scenario: it is often impossible for rankers to specify full rankings over a very large number of items making the use of partial rankings a necessity. Different rankers may choose to output partial rankings of different types, for example, one ranker can output $3 \prec \{1,2\}$ (3 is preferred to both 1 and 2) and another ranker can output $\{1,3\} \prec 2$ (both 1 and 3 are preferred to 2). By considering the asymptotics $n \to \infty$ in addition to $m \to \infty$ (m being the number of samples) we provide a more realistic analysis for a large (and potentially growing) number of items. Computational feasibility is a major concern since most ranking models are incapable of modeling the data sets mentioned above due to their large n.

We continue next by reviewing basic concepts concerning partially ranked data and the Mallows model, and then proceed to define our non-parametric estimator. We conclude by demonstrating computational efficiency, statistical properties, and some experiments.

2. Permutations and Cosets

We begin by reviewing some basic concepts concerning permutations, with some of the notations and definitions borrowed from Critchlow (1985).

A permutation π is a bijective function $\pi : \{1,\ldots,n\} \to \{1,\ldots,n\}$ associating with each item $i \in \{1,\ldots,n\}$ a rank $\pi(i) \in \{1,\ldots,n\}$. In other words, $\pi(i)$ denotes the rank given to item i and $\pi^{-1}(i)$ denotes the item assigned to rank i. We denote a permutation π using the following vertical bar notation $\pi^{-1}(1)|\pi^{-1}(2)|\cdots|\pi^{-1}(n)$. For example, the permutation $\pi(1) = 2, \pi(2) = 3, \pi(3) = 1$ would be denoted as $3|1|2$. In this notation the numbers correspond to items and the locations of the items in their corresponding compartments correspond to their ranks. The collection of all permutations of n items forms the non-Abelian symmetric group of order n, denoted by $\mathfrak{S}_n$, using function composition as the group operation $\pi\sigma = \pi \circ \sigma$. We denote the identity permutation by e.

The concept of inversions and the result below will be of great use later on.

Definition 1 *The inversion set of a permutation π is the set of pairs*

$$U(\pi) \stackrel{\text{def}}{=} \{(i,j) : i < j,\ \pi(i) > \pi(j)\} \subset \{1,\ldots,n\} \times \{1,\ldots,n\}$$

whose cardinality is denoted by $i(\pi) \stackrel{\text{def}}{=} |U(\pi)|$.

For example, $i(e) = |\emptyset| = 0$, and $i(3|2|1|4) = |\{(1,2),(1,3),(2,3)\}| = 3$.

Proposition 1 (for example, Stanley, 2000) *The map $\pi \mapsto U(\pi)$ is a bijection.*

When n is large, the enormous number of permutations raises difficulties in using the symmetric group for modeling rankings. A reasonable solution is achieved by considering partial rankings which correspond to cosets of the symmetric group. For example, the subgroup of $\mathfrak{S}_n$ consisting of

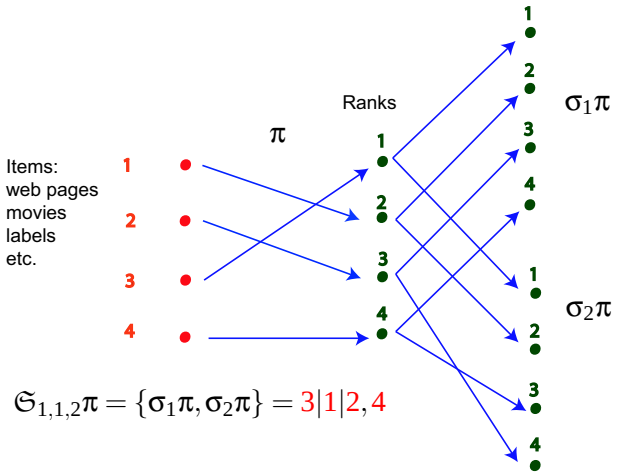

Figure 1: A partial ranking corresponds to a coset or a set or permutations

all permutations that fix the top k positions is denoted $\mathfrak{S}_{1,\ldots,1,n-k} = \{\pi \in \mathfrak{S}_n : \pi(i) = i,\ i = 1,\ldots,k\}$. The right coset $\mathfrak{S}_{1,\ldots,1,n-k}\pi = \{\sigma\pi : \sigma \in \mathfrak{S}_{1,\ldots,1,n-k}\}$ is the set of permutations consistent with the ordering of π on the k top-ranked items. It may thus be interpreted as a partial ranking of the top k items, that does not contain any information concerning the relative ranking of the bottom $n-k$ items. The set of all such partial rankings forms the quotient space $\mathfrak{S}_n/\mathfrak{S}_{1,\ldots,1,n-k}$. Figure 1 illustrates the identification of a coset as a partial ranking of the top 2 out of 4 items.

We generalize the above relationship between partial rankings and cosets through the following definition of a composition.

Definition 2 *A composition of n is a sequence $\gamma = (\gamma_1,\ldots,\gamma_r)$ of positive integers whose sum is n.*

Note that in contrast to a partition, in a composition the order of the integers matters. A composition $\gamma = (\gamma_1,\ldots,\gamma_r)$ corresponds to a partial ranking with γ_1 items in the first position, γ_2 items in the second position and so on. For such a partial ranking it is known that the first set of γ_1 items are to be ranked before the second set of γ_2 items etc., but no further information is conveyed about the orderings within each set. The partial ranking introduced earlier $\mathfrak{S}_{1,\ldots,1,n-k}\pi$ of the top k items is a special case corresponding to $\gamma = (1,\ldots,1,n-k)$.

More formally, let $N_1 = \{1,\ldots,\gamma_1\}, N_2 = \{\gamma_1+1,\ldots,\gamma_1+\gamma_2\},\cdots,N_r = \{\gamma_1+\cdots+\gamma_{r-1}+1,\ldots,n\}$. The subgroup $\mathfrak{S}_\gamma$ is defined as the set of all permutations $\pi \in \mathfrak{S}_n$ for which the following set equalities hold (the two sets on the left hand side and right hand side of the equality contain the same elements)

$$\pi(N_i) = N_i \qquad i = 1,\ldots,r.$$

In other words, the subgroup $\mathfrak{S}_\gamma$ contains permutations that only permute within each set N_i. It can be shown that the subgroup $\mathfrak{S}_\gamma$ is isomorphic to the product of subgroups $\mathfrak{S}_{\gamma_1} \times \cdots \times \mathfrak{S}_{\gamma_r}$ and is sometimes described by that product for notational purposes. A partial ranking of type γ is equivalent to a coset $\mathfrak{S}_\gamma\pi = \{\sigma\pi : \sigma \in \mathfrak{S}_\gamma\}$ and the set of such partial rankings forms the quotient space $\mathfrak{S}_n/\mathfrak{S}_\gamma$.

The vertical bar notation described above for permutations is particularly convenient for denoting partial rankings. We list items $1,\ldots,n$ separated by vertical bars, indicating that items on the left side of each vertical bar are preferred to (ranked higher than) items on the right side of the bar. On the other hand, there is no knowledge concerning the preference of items that are not separated by one or more vertical bars. For example, the partial ranking displayed in Figure 1 is denoted by $3|1|2,4$. The ordering of items not separated by a vertical line is meaningless, and for consistency we use the conventional ordering, for example, $1|2,3|4$ rather than the equivalent $1|3,2|4$.

The set of all partial rankings

$$\mathfrak{W}_n \stackrel{\text{def}}{=} \{\mathfrak{S}_\gamma \pi : \pi \in \mathfrak{S}_n, \forall \gamma\} \tag{2}$$

which includes the set of full rankings $\mathfrak{S}_n$, is a subset of all possible partial orders on $\{1,\ldots,n\}$. While the formalism of partial rankings in $\mathfrak{W}_n$ cannot realize all partial orderings, it is sufficiently powerful to include many useful and naturally occurring orderings as special cases. Furthermore, as demonstrated in later sections, it enables simplification of the otherwise overwhelming computational difficulty. Special cases of particular interest are the following partial rankings

- $\pi \in \mathfrak{S}_n$ corresponds to a permutation or a full ordering, for example, $3|2|4|1$.
- $\mathfrak{S}_{1,n-1}\pi$, for example, $3|1,2,4$, corresponds to selection of the top alternative. An example for such a ranking is a classification of x by a response variable $y \in \mathcal{Y} = \{1,\ldots,n\}$.
- $\mathfrak{S}_{1,\ldots,1,n-k}\pi$, for example, $1|3|2,4$, corresponds to full ordering of the top k items. An example for such a ranking is a ranked list of the top k webpages output by search engines in response to a query.
- $\mathfrak{S}_{k,n-k}\pi$, for example, $1,2,4|3,5$, corresponds to a more preferred and a less preferred dichotomy. Alternatively the dichotomy can be interpreted as right and wrong or relevant and irrelevant. An example for such a ranking is classification of alternatives into desirable and undesirable.
- $\mathfrak{S}_{1,\ldots,1,n-k-t,1,\ldots,1}\pi$, for example, $5|1|2,4,7|6|3|8$, corresponds to full ordering of the top k and the bottom t items. An example for such a ranking is a list of the safest and the most dangerous U.S. cities.[6]
- $\mathfrak{S}_{k,n-k-t,t}\pi$, for example, $1,5|2,4,7|3,6,8$, corresponds to a trichotomy of items. An example for such a ranking is selection of preferred and non-preferred items from a list.

Traditionally, data from each one of the special cases above was modeled using different tools and was considered fundamentally different. That problem was aggravated as different special cases were usually handled by different communities (statistics, computer science, information retrieval). As a first step towards presenting a unified framework for modeling partially ranked data, Lebanon and Lafferty (2003) demonstrated equivalence between several popular conditional models. We continue along this line and present in this paper a non-parametric framework capable of efficiently modeling a large variety of partially ranked data.

In constructing a statistical model on permutations or cosets, it is essential to relate one permutation to another. We do this using a distance function on permutations $d : \mathfrak{S}_n \times \mathfrak{S}_n \to \mathbb{R}$ that

6. List can be found at `http://www.infoplease.com/ipa/A0921299.html`.

satisfies the usual metric function properties, and in addition is invariant under right action of the symmetric group (Critchlow, 1985)

$$d(\pi,\sigma) = d(\pi\tau,\sigma\tau) \quad \forall\, \pi,\sigma,\tau \in \mathfrak{S}_n. \tag{3}$$

The invariance requirement (3) ensures that the distance does not change if the labeling of the items $\{1,\ldots,n\}$ (which is assumed to be arbitrary) is permuted.

There have been many propositions for such right-invariant distance functions, the most popular of them being Kendall's tau (Kendall, 1938)

$$d(\pi,\sigma) = \sum_{i=1}^{n-1}\sum_{l>i} I(\pi\sigma^{-1}(i) - \pi\sigma^{-1}(l)) \tag{4}$$

where $I(x) = 1$ for $x > 0$ and $I(x) = 0$ otherwise. Kendall's tau $d(\pi,\sigma)$ (4) measures the number of pairs of items for which π and σ have opposing orderings (also called disconcordant pairs). An equivalent definition for Kendall's tau is the minimum number of adjacent transpositions needed to bring π^{-1} to σ^{-1} (adjacent transposition flips a pair of items having adjacent ranks). By right invariance, $d(\pi,\sigma) = d(\pi\sigma^{-1}, e)$ which, for Kendall's tau equals the number of inversions $i(\pi\sigma^{-1})$. This is an important observation that will allow us to simplify many expressions concerning Kendall's tau using the combinatorial properties of inversions.

Kendall's tau $d(\pi,\sigma), \pi,\sigma \in \mathfrak{S}_n$ takes values between 0 for $\pi = \sigma$ and $n(n-1)/2$. It is sometimes desirable to consider the normalized Kendall's tau

$$d_n(\pi,\sigma) = \frac{2}{n(n-1)} \sum_{i=1}^{n-1}\sum_{l>i} I(\pi\sigma^{-1}(i) - \pi\sigma^{-1}(l)) \tag{5}$$

whose range is $[0,1]$ and consequentially may be compared across different values of n.

3. The Mallows Model and its Extension to Partial Rankings

The Mallows model (Mallows, 1957) is a location-scale model on permutations based on Kendall's tau distance

$$p_\kappa(\pi) = \exp\left(-c\, d(\pi,\kappa) - \log\psi(c)\right) \qquad \pi,\kappa \in \mathfrak{S}_n \quad c \in \mathbb{R}_+.$$

The normalization term $\psi(c) = \sum_{\pi\in\mathfrak{S}_n} \exp(-c\, d(\pi,\kappa))$ does not depend on the location parameter κ and has the closed form

$$\begin{aligned}\psi(c) &= \sum_{\pi\in\mathfrak{S}_n} e^{-c\, d(\pi,\kappa)} \\ &= (1+e^{-c})(1+e^{-c}+e^{-2c})\cdots(1+e^{-c}+\cdots+e^{-(n-1)c}) \\ &= \prod_{j=1}^{n} \frac{1-e^{-jc}}{1-e^{-c}}\end{aligned} \tag{6}$$

as shown by the fact that $d(\pi,\kappa) = i(\pi\kappa^{-1})$ and the following proposition.

Proposition 2 (for example, Stanley, 2000) *For $q > 0$, $\sum_{\pi\in\mathfrak{S}_n} q^{i(\pi)} = \prod_{j=1}^{n-1}\sum_{k=0}^{j} q^k$.*

Proof Due to the bijection between permutations and sets of inversions expressed in Proposition 1

$$\sum_{\pi\in\mathfrak{S}_n} q^{i(\pi)} = \sum_{a_1=0}^{n-1}\sum_{a_2=0}^{n-2}\cdots\sum_{a_n=0}^{0} q^{a_1+\ldots+a_n} = \left(\sum_{a_1=0}^{n-1} q^{a_1}\right)\left(\sum_{a_2=0}^{n-2} q^{a_2}\right)\cdots\left(\sum_{a_n=0}^{0} q^{a_n}\right)$$
$$= (1+q+\cdots+q^{n-1})\cdots(1+q+q^2)(1+q)1.$$

■

The Mallows model has been motivated on axiomatic grounds by Mallows and has been a major focus of statistical modeling on permutations. Various extensions of the Mallows model may be found in Fligner and Verducci (1986, 1988, 1993). One particular extension to partial rankings is to consider a partial ranking as censored data equivalent to the set of permutations in its related coset. In other words, we define the probability the model assigns to the partial ranking $\mathfrak{S}_\gamma\pi$ by

$$\sum_{\tau\in\mathfrak{S}_\gamma\pi} p_\kappa(\tau) = \psi^{-1}(c)\sum_{\tau\in\mathfrak{S}_\gamma\pi}\exp\left(-c\,d(\tau,\kappa)\right). \tag{7}$$

Fligner and Verducci (1986) showed that in the case of $\gamma = (1,\ldots,1,n-k)$ the summation in (7) has a simple closed form. However, the apparent absence of a closed form formula for more general partial rankings prevented the widespread use of Equation 7 for large n and encouraged more ad-hoc and heuristic models (Critchlow, 1985; Marden, 1996). Section 7 describes an efficient computational procedure for computing (7) for more general partial ranking types γ.

4. The Ranking Lattice

Partial rankings $\mathfrak{S}_\gamma\pi$ relate to each other in a natural way by expressing more general, more specific or inconsistent ordering. We define below the concepts of partially ordered sets and lattices and then relate them to partial rankings by considering the set of partial rankings $\mathfrak{W}_n$ as a lattice. Some of the definitions below are taken from Stanley (2000), where a thorough introduction to posets can be found.

Definition 3 *A partially ordered set or poset $(Q,\preceq)$, is a set Q endowed with a binary relation $\preceq$ satisfying $\forall x,y,z\in Q$ (i) reflexibility: $x\preceq x$, (ii) anti-symmetry: $x\preceq y$ and $y\preceq x\Rightarrow x=y$, and (iii) transitivity: $x\preceq y$ and $y\preceq z\Rightarrow x\preceq z$.*

We write $x\prec y$ when $x\preceq y$ and $x\neq y$. We say that y covers x when $x\prec y$ and there is no $z\in Q$ such that $x\prec z\prec y$. A finite poset is completely described by its covering relation. The planar Hasse diagram of $(Q,\preceq)$ is the graph connecting the elements of Q as nodes using edges that correspond to the covering relation. An additional requirement is that if y covers x then y is drawn higher than x. Two elements x,y are comparable if $x\preceq y$ or $y\preceq x$ and otherwise are incomparable. The set of partial rankings $\mathfrak{W}_n$ defined in (2) is naturally endowed with the partial order of ranking refinement, that is, $\pi\prec\sigma$ if π refines σ or alternatively if we can get from π to σ by dropping vertical lines (Lebanon and Lafferty, 2003). Figure 2 shows the Hasse diagram of $\mathfrak{W}_3$ and a partial Hasse diagram of $\mathfrak{W}_4$.

An interesting visualization of Kendall's tau distance $d(\pi,\sigma),\pi,\sigma\in\mathfrak{S}_n$ in terms of the Hasse diagram is that it is the minimum number of up and down moves needed to get from π to σ on the

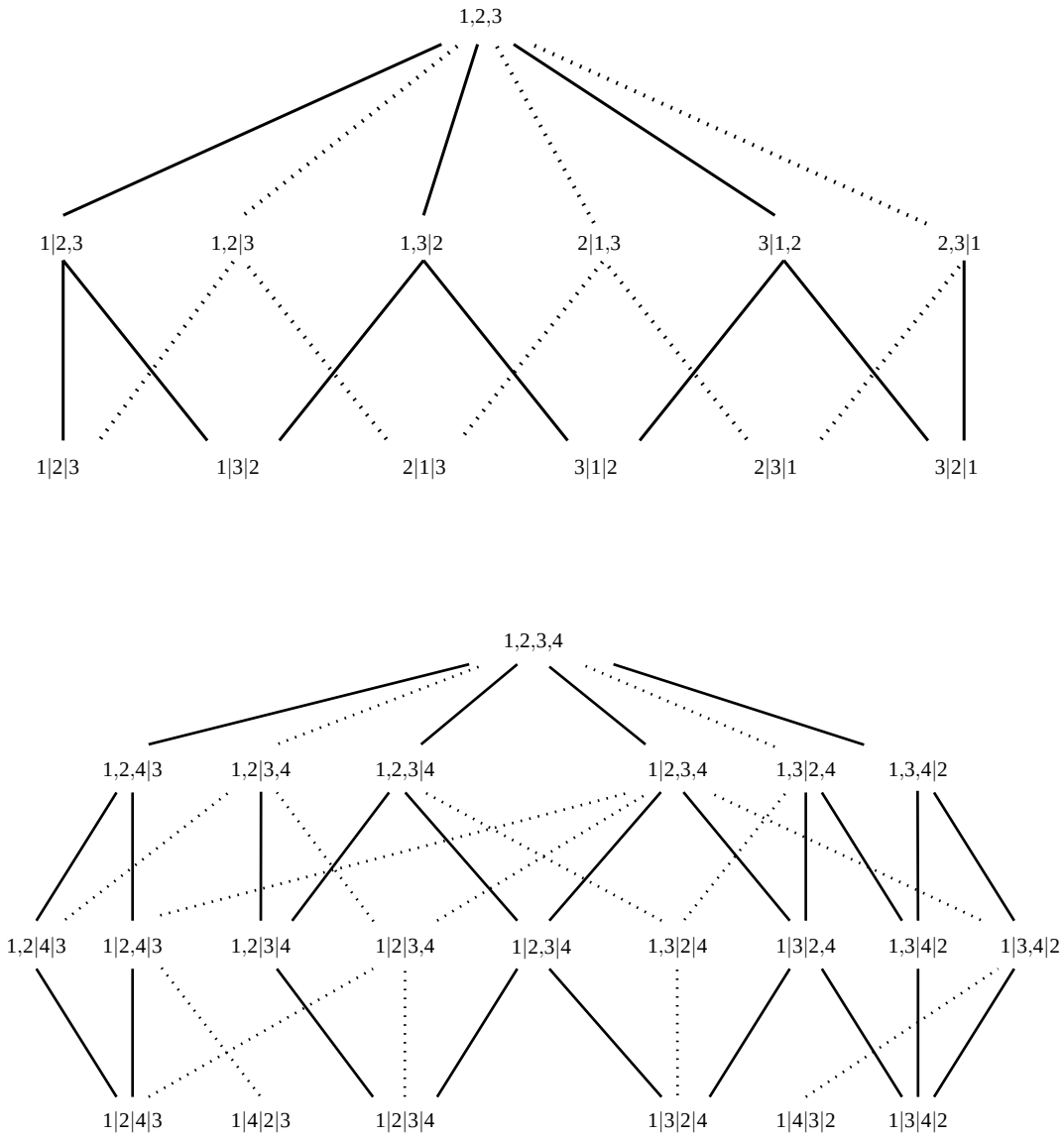

Figure 2: The Hasse diagram of $\mathfrak{W}_3$ (top) and a partial Hasse diagram of $\mathfrak{W}_4$ (bottom). Some of the lines are dotted for 3D visualization purposes (think 3D).

Hasse diagram. For example, in Figure 2 (top) we have $d(1|2|3,3|2|1) = 3$ realized by the three up-down moves along the shortest path

$$1|2|3\ (\nearrow 1,2|3 \searrow 2|1|3)\ (\nearrow 2|1,3 \searrow 2|3|1)\ (\nearrow 2,3|1 \searrow 3|2|1).$$

A lower bound z of two elements in a poset x,y satisfies $z \preceq x$ and $z \preceq y$. The greatest lower bound of x,y or infimum is a lower bound of x,y that is greater than or equal to any other lower bound of x,y. Infimum, and the analogous concept of supremum are denoted by $x \wedge y$ and $x \vee y$ or

$\bigwedge\{x_1,\ldots,x_k\}$ and $\bigvee\{x_1,\ldots,x_k\}$ respectively. Two elements $x,y \in \mathfrak{W}_n$ are said to be consistent if there exists a lower bound in $\mathfrak{W}_n$. Note that consistency is a weaker relation than comparability. For example, $1|2,3|4$ and $1,2|3,4$ are consistent but incomparable while $1|2,3|4$ and $2|1,3|4$ are both inconsistent and incomparable. Using the vertical bar notation, two elements are inconsistent iff there exist two items i,j that appear on opposing sides of a vertical bar in x and y, that is, $x = \cdots i|j \cdots$ while $y = \cdots j|i \cdots$. A poset for which $\wedge$ and $\vee$ always exist is called a lattice. Lattices satisfy many useful combinatorial properties - one of which is that they are completely described by the $\wedge$ and $\vee$ operations. In fact lattices are often defined by the supremum and infimum relation, rather than by the partial order. While the ranking poset is not a lattice, it may be turned into one by augmenting it with a minimum element $\hat{0}$.

Proposition 3 *The union $\tilde{\mathfrak{W}}_n \stackrel{\text{def}}{=} \mathfrak{W}_n \cup \{\hat{0}\}$ of the ranking poset and a minimum element is a lattice.*

Proof Since $\tilde{\mathfrak{W}}_n$ is finite, it is enough to show existence of $\wedge,\vee$ for pairs of elements (Stanley, 2000). We begin by showing existence of $x \wedge y$. If x,y are inconsistent, there is no lower bound in $\mathfrak{W}_n$ and therefore the unique lower bound $\hat{0}$ is also the infimum $x \wedge y$. If x,y are consistent, their infimum may be obtained as follows. Since x and y are consistent, we do not have a pair of items i,j appearing as $i|j$ in x and $j|i$ in y. As a result we can form a lower bound z to x,y by starting with a list of numbers and adding the vertical bars that are in either x or y, for example for $x = 3|1,2,5|4$ and $y = 3|2|1,4,5$ we have $z = 3|2|1,5|4$. The resulting $z \in \mathfrak{W}_n$, is smaller than x and y since by construction it contains all preferences (encoded by vertical bars) in x and y. It remains to show that for every other lower bound z' of x and y we have $z' \preceq z$. If z' is comparable to z, $z' \preceq z$ since removing any vertical bar from z results in an element that is not a lower bound. If z' is not comparable to z, then both z,z' contain the vertical bars in x and vertical bars in y possibly with some additional ones. By construction z contains only the essential vertical bars to make it a lower bound and hence $z' \prec z$, contradicting the assumption that z,z' are non-comparable.

By Proposition 3.3.1 of Stanley (2000) a poset for which an infimum is always defined and that has a supremum element is necessarily a lattice. Since we just proved that $\wedge$ always exists for $\tilde{\mathfrak{W}}_n$ and $1,\ldots,n = \bigvee \tilde{\mathfrak{W}}_n$, the proof is complete. ■

5. Probabilistic Models on the Ranking Lattice

The ranking lattice is a convenient framework to define and study probabilistic models on partial rankings. Given a probability model p on $\mathfrak{S}_n$, we define the functions $h,g : \tilde{\mathfrak{W}}_n \to [0,1]$

$$h(\alpha) = \begin{cases} p(\alpha) & \alpha \in \mathfrak{S}_n \\ 0 & \alpha \in \tilde{\mathfrak{W}}_n \setminus \mathfrak{S}_n \end{cases}$$
$$g(\alpha) = \sum_{\beta \in \tilde{\mathfrak{W}}_n : \beta \preceq \alpha} h(\beta). \tag{8}$$

Interpreting partial rankings $\mathfrak{S}_\gamma\pi \in \tilde{\mathfrak{W}}_n$ as the disjoint union of the events defined by the coset $\mathfrak{S}_\gamma\pi$ we have that

$$g(\mathfrak{S}_\gamma\pi) = \sum_{\tau \in \mathfrak{S}_\gamma\pi} p(\tau) \tag{9}$$

may be interpreted as the probability under p of the disjoint union $\mathfrak{S}_\gamma\pi$ of permutations. We refer to the function g as the partial ranking or lattice version of p. The motivation for defining g through h and not directly through p is that Equation (8) may be described and computed by the mechanism of Möbius inversion on lattices. More specifically, the Möbius inversion on lattices states that for two arbitrary real-valued functions on a lattice $h, g : \tilde{\mathfrak{W}}_n \to [0,1]$ we have

$$g(\tau) = \sum_{\tau' \preceq \tau} h(\tau') \quad \text{iff} \quad h(\tau) = \sum_{\tau' \preceq \tau} g(\tau')\mu(\tau', \tau) \qquad \tau, \tau \in \tilde{\mathfrak{W}}_n$$

where $\mu : \tilde{\mathfrak{W}}_n \times \tilde{\mathfrak{W}}_n \to \mathbb{R}$ is the Möbius function of the lattice $\tilde{\mathfrak{W}}_n$. In a certain sense this relationship between p and g generalizes the relationship between a probability mass function and the corresponding cdf. More details on Möbius functions and Möbius inversion on lattices and their computation may be found in Stanley (2000).

The function g is defined on the entire lattice, but when restricted to partial rankings of the same type $G = \{\mathfrak{S}_\gamma\pi : \pi \in \mathfrak{S}_n\} \subset \tilde{\mathfrak{W}}_n$, constitutes a normalized probability distribution on G. Estimating and examining a restriction of g to a subset $H \subset \tilde{\mathfrak{W}}_n$ (note that in general H may include more than one coset space G) rather than the function p is particularly convenient in cases of large n since H is often much smaller than the unwieldy $\mathfrak{S}_n$. In such cases it is tempting to specify the function g directly on H without referring to an underlying permutation model. However, doing so may lead to probabilistic contradictions such as $g(\mathfrak{S}_\gamma\pi) < g(\mathfrak{S}_\lambda\sigma)$ for $\mathfrak{S}_\lambda\sigma \subset \mathfrak{S}_\gamma\pi$. To avoid these and other probabilistic contradictions, g needs to satisfy a set of linear constraints equivalent to the existence of an underlying permutation model. Figure 3 illustrates this problem for partial rankings with the same (left) and different (right) number of vertical bars. A simple way to avoid such contradictions and satisfy the constraints is to define g indirectly in terms of a permutation model p as in (9). Applied to the context of statistical estimation, we define the estimator $\hat{g}$ in terms of an estimator $\hat{p}$ of the underlying permutation model p.

In addition to this construction which logically occurs after obtaining the estimator $\hat{p}$, we also need to consider how to use partially ranked data in the process of obtaining the estimator $\hat{p}$. Fully ranked data is often not available for large n since it is difficult for rankers (both human and others) to express with confidence full orderings over many items. Instead, the inference needs to be conducted based on a set of partial rankings

$$D = \{\mathfrak{S}_{\gamma_i}\pi_i : i = 1, \ldots, m\}. \tag{10}$$

A general way of using D in (10) to estimate $\hat{p}$ both parametrically and non-parametrically is to consider partially ranked data as censored or missing data. In other words, in the process of estimating $\hat{p}$, the data $\mathfrak{S}_\gamma\pi$ is considered as a single unknown permutation $\sigma \in \mathfrak{S}_\gamma\pi$ that is lost through a censoring process. Assuming uniformly random censoring in a parametric setting, we obtain the following observed likelihood with respect to the partially ranked data set D

$$\ell(\theta|D) = \sum_{i=1}^{m} \log \frac{1}{|\mathfrak{S}_{\gamma_i\pi_i}|} \sum_{\sigma \in \mathfrak{S}_{\gamma_i}\pi_i} p_\theta(\sigma) = \sum_{i=1}^{m} \log \sum_{\sigma \in \mathfrak{S}_{\gamma_i}\pi_i} p_\theta(\sigma) + \text{const.}$$

While the above likelihood function can be efficiently computed using tools developed in Section 7, its maximization is extremely difficult due to the discrete nature of the parametric space. In the next section we explore in detail a non-parametric kernel smoothing alternative to estimating p and g based on partially ranked data.

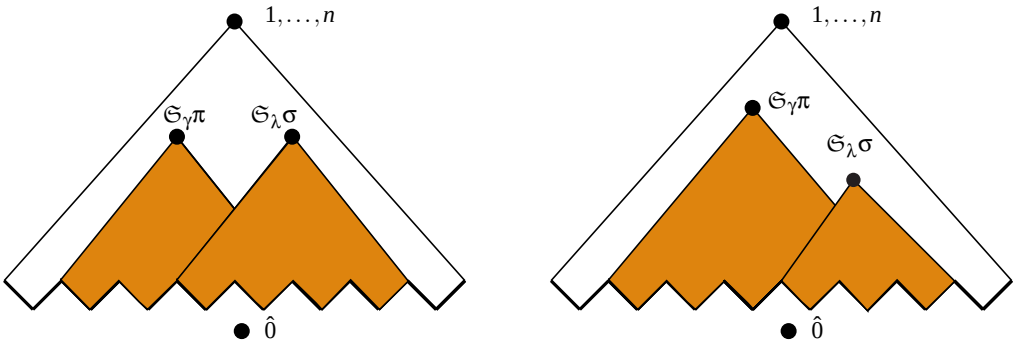

Figure 3: Two partial rankings with the same (left) and different (right) number of vertical bars in the Hasse diagram of $\tilde{\mathfrak{W}}_n$. The big triangles are schematic illustration of the Hasse diagram of $\mathfrak{W}_n$ as displayed in Figure 2 (top) with permutations occupying the bottom level illustrated by the jagged line. The shaded regions correspond to order-intervals, that is, all elements smaller or equal to the top vertices which correspond to partial rankings. To avoid probabilistic contradictions, the values of g at two non-disjoint partial rankings $\mathfrak{S}_\gamma\pi, \mathfrak{S}_\lambda\sigma$ cannot be specified in an independent manner.

6. Non-Parametric Kernel Smoothing on Partial Rankings

The Mallows model, which at first glance appears as a simple and effective analogue of the Gaussian distribution, suffers from several drawbacks. Its unimodal assumption is often too restrictive for high n as well as for low n (see experiments in Section 9). Another major drawback is that the location parameter space $\mathfrak{S}_n$ is discrete, making the maximum likelihood procedure an impossibly large discrete search problem.

The unimodality and symmetry of the Mallows model make it a good choice for use as a kernel in non-parametric smoothing. Since the normalization term ψ does not depend on the location parameter (6), the kernel smoothing estimator for p is

$$\hat{p}(\pi) = \frac{1}{m\,\psi(c)} \sum_{i=1}^{m} \exp(-c\,d(\pi,\pi_i)) \quad \pi,\pi_i \in \mathfrak{S}_n \tag{11}$$

assuming the data consists of complete rankings $\pi_1,\ldots,\pi_m \sim p$. Note that the kernel parameter c acts as an inverse scale parameter whose role is similar but inversely related to the traditional bandwidth parameter h in kernel smoothing (Wand and Jones, 1995).

In case the available data is partially ranked $D = \{\mathfrak{S}_{\gamma_i}\pi_i : i = 1,\ldots,m\}$ and obtained by uniform censoring as described in the previous section the kernel smoothing estimator becomes

$$\hat{p}(\pi) = \frac{1}{m\,\psi(c)} \sum_{i=1}^{m} \frac{1}{|\mathfrak{S}_{\gamma_i}|} \sum_{\tau\in\mathfrak{S}_{\gamma_i}\pi_i} \exp(-c\,d(\pi,\tau)) \qquad \pi \in \mathfrak{S}_n \tag{12}$$

where we used the fact that $|\mathfrak{S}_{\gamma_i}\pi_i| = |\mathfrak{S}_{\gamma_i}e| = |\mathfrak{S}_{\gamma_i}|$. The lattice or partial ranking version $\hat{g}$ corresponding to $\hat{p}$ in (12) is

$$\hat{g}(\mathfrak{S}_\lambda \pi) = \frac{1}{m\,\psi(c)} \sum_{i=1}^{m} \frac{1}{|\mathfrak{S}_{\gamma_i}|} \sum_{\kappa \in \mathfrak{S}_\lambda \pi} \sum_{\tau \in \mathfrak{S}_{\gamma_i}\pi_i} \exp(-c\,d(\kappa,\tau)) \qquad \mathfrak{S}_\gamma \pi \in \tilde{\mathfrak{W}}_n. \tag{13}$$

In the next section we derive efficient calculations and in some cases closed forms for expressions (12)-(13). These calculations are efficient even for large n as their complexities depend on the complexity of the compositions λ and $\gamma_1,\ldots,\gamma_m$ rather than on $n!$ or even n. We then move on to explore the bias and variance of $\hat{p}$ in Section 8 and describe practical applications of $\hat{p},\hat{g}$ and some experiments.

7. Efficient Computation and Inversion Combinatorics

In order to apply the estimators $\hat{p},\hat{g}$ in practice, it is crucial that the inner summations in Equations (12)-(13) be computed efficiently. We can achieve efficient computation of these summations by considering how the pairs constituting inversions $i(\tau)$ decompose with respect to certain cosets.

Proposition 4 *The following decomposition of $i(\tau)$ with respect to a composition $\gamma = (\gamma_1,\ldots,\gamma_r)$ holds*

$$i(\tau) = \sum_{k=1}^{r} a_k^\gamma(\tau) + \sum_{k=1}^{r}\sum_{l=k+1}^{r} b_{kl}^\gamma(\tau) \qquad \forall \tau \in \mathfrak{S}_n \tag{14}$$

where

$$a_k^\gamma(\tau) \stackrel{\text{def}}{=} \left| \left\{ (s,t) : s<t\,, \sum_{j=1}^{k-1}\gamma_j < \tau(t) < \tau(s) \le \sum_{j=1}^{k}\gamma_j \right\} \right|$$

$$b_{kl}^\gamma(\tau) \stackrel{\text{def}}{=} \left| \left\{ (s,t) : s<t\,, \sum_{j=1}^{k-1}\gamma_j < \tau(t) \le \sum_{j=1}^{k}\gamma_j \le \sum_{j=1}^{l-1}\gamma_j < \tau(s) \le \sum_{j=1}^{l}\gamma_j \right\} \right|.$$

Proof The set appearing in the definition of $a_k^\gamma(\tau)$ contains all pairs (s,t) that are inversions of τ and whose ranks appear in the k-compartment of the composition γ. The set appearing in the definition of $b_{kl}^\gamma(\tau)$ contains pairs (s,t) that are inversions of τ and for which s and t appear in the l and k compartments of γ respectively. Since any inversion pair appears in either one or two compartments, the above forms a partition of the inversion set. The decomposition holds since $i(\tau)$, the cardinality of the inversion set of the permutation τ, equals the summation of the cardinality of each subset in the partition. ■

Equation (14) actually represents a family of decompositions as it holds for all possible compositions γ. For example, $i(\tau) = 4$ for $\tau = 4|1|3|2$, with inversions $(1,4),(2,4),(3,4),(2,3)$ for τ. For the composition $\gamma = (2,2)$, the first compartment contains the inversion $(1,4)$ and so $a_1^\gamma(\tau) = 1$. The second compartment contains the inversion $(2,3)$ and so $a_2^\gamma(\tau) = 1$. The cross compartment inversions are $(2,4),(3,4)$ making $b_{12}^\gamma(\tau) = 2$.

The significance of (14) is that as we sum over all representatives of the coset $\tau \in \mathfrak{S}_\gamma\pi$ the cross compartmental inversions $b_{kl}^\gamma(\tau)$ remain constant while the within-compartmental inversions

$a_k^\gamma(\tau)$ vary over all possible combinations. As a result we obtain the following generalization of Proposition 2.

Proposition 5 *For $\pi \in \mathfrak{S}_n$, $q > 0$, and a composition $\gamma = (\gamma_1, \ldots, \gamma_r)$ we have*

$$\sum_{\tau \in \mathfrak{S}_\gamma \pi} q^{i(\tau)} = q^{\sum_{k=1}^r \sum_{l=k+1}^r b_{kl}^\gamma(\pi)} \prod_{s=1}^r \prod_{j=1}^{\gamma_s - 1} \sum_{k=0}^j q^k. \tag{15}$$

Proof

$$\begin{aligned}
\sum_{\tau \in \mathfrak{S}_\gamma \pi} q^{i(\tau)} &= \sum_{\tau \in \mathfrak{S}_\gamma \pi} q^{\sum_{k=1}^r a_k^\gamma(\tau) + \sum_{k=1}^r \sum_{l=k+1}^r b_{kl}^\gamma(\tau)} \\
&= q^{\sum_{k=1}^r \sum_{l=k+1}^r b_{kl}^\gamma(\pi)} \sum_{\tau \in \mathfrak{S}_\gamma \pi} q^{\sum_{k=1}^r a_k^\gamma(\tau)} \\
&= q^{\sum_{k=1}^r \sum_{l=k+1}^r b_{kl}^\gamma(\pi)} \prod_{s=1}^r \sum_{\tau \in \mathfrak{S}_{\gamma_s}} q^{i(\tau)} \\
&= q^{\sum_{k=1}^r \sum_{l=k+1}^r b_{kl}^\gamma(\pi)} \prod_{s=1}^r \prod_{j=1}^{\gamma_s - 1} \sum_{k=0}^j q^k.
\end{aligned}$$

Above, we used two ideas: (i) disconcordant pairs between two different compartments of the coset $\mathfrak{S}_\gamma \pi$ are invariant under change of the coset representative, and (ii) the number of disconcordant pairs within a compartment varies over all possible choices enabling the replacement of the summation by a sum over a lower order symmetric group. ■

An important feature of (15) is that only the first and relatively simple term $q^{\sum_{k=1}^r \sum_{l=k+1}^r b_{kl}^\gamma(\pi)}$ depends on π. The remaining terms depend only on the partial ranking type γ and thus may be pre-computed and tabulated for efficient computation.

Corollary 1

$$\sum_{\tau \in \mathfrak{S}_\gamma \pi} q^{i(\tau\kappa)} = q^{\sum_{k=1}^r \sum_{l=k+1}^r b_{kl}^\gamma(\pi\kappa)} \prod_{s=1}^r \prod_{j=1}^{\gamma_s - 1} \sum_{k=0}^j q^k \qquad \kappa \in \mathfrak{S}_n.$$

Proof Using group theory, it can be shown that the set equality $(\mathfrak{S}_\gamma \pi)\kappa = \mathfrak{S}_\gamma(\pi\kappa)$ holds. As a result, $\sum_{\tau \in \mathfrak{S}_\gamma \pi} q^{i(\tau\kappa)} = \sum_{\tau' \in \mathfrak{S}_\gamma(\pi\kappa)} q^{i(\tau')}$. Proposition 5 completes the proof. ■

Corollary 2 *The partial ranking version g corresponding to the Mallows kernel p_κ is*

$$\begin{aligned}
p_\kappa(\mathfrak{S}_\gamma \pi) &= \frac{\prod_{s=1}^r \prod_{j=1}^{\gamma_s - 1} \sum_{k=0}^j e^{-kc}}{\prod_{j=1}^{n-1} \sum_{k=0}^j e^{-kc}} \, e^{-c \sum_{k=1}^r \sum_{l=k+1}^r b_{kl}^\gamma(\pi\kappa^{-1})} \\
&\propto e^{-c \sum_{k=1}^r \sum_{l=k+1}^r b_{kl}^\gamma(\pi\kappa^{-1})}.
\end{aligned}$$

Proof Using Corollary 1 we have

$$\begin{aligned} g(\mathfrak{S}_\gamma\pi) &= \sum_{\tau\in\mathfrak{S}_\gamma\pi} p_\kappa(\tau) = \frac{\sum_{\tau\in\mathfrak{S}_\gamma\pi}\exp(-c\,d(\tau,\kappa))}{\sum_{\tau\in\mathfrak{S}_n}\exp(-c\,d(\tau,\kappa))} \\ &= \frac{\sum_{\tau\in\mathfrak{S}_\gamma\pi}\exp(-c\,i(\tau\kappa^{-1}))}{\prod_{j=1}^{n-1}\sum_{k=0}^{j}e^{-kc}} = \frac{\sum_{\tau\in\mathfrak{S}_\gamma\pi}(\exp(-c))^{i(\tau\kappa^{-1})}}{\prod_{j=1}^{n-1}\sum_{k=0}^{j}e^{-kc}} \\ &= e^{-c\sum_{k=1}^{r}\sum_{l=k+1}^{r}b_{kl}^{\gamma}(\pi\kappa^{-1})}\frac{\prod_{s=1}^{r}\prod_{j=1}^{\gamma_s-1}\sum_{k=0}^{j}e^{-kc}}{\prod_{j=1}^{n-1}\sum_{k=0}^{j}e^{-kc}}. \end{aligned}$$

■

Despite its daunting appearance, the expression in Corollary 2 can be computed relatively easily. The fraction does not depend on π or κ and in fact may be considered as a normalization constant that may be easily pre-computed and tabulated. The remaining term is relatively simple and depends on the location parameter κ and the coset representative π. Corollary 2 and Proposition 6 below, provide efficient computation for the estimators (12), (13).

Proposition 6

$$\sum_{\sigma\in\mathfrak{S}_\lambda\pi_1}\sum_{\tau\in\mathfrak{S}_\gamma\pi_2}e^{-c\,d(\sigma,\tau)} = \left(\sum_{\tau\in\pi_1\pi_2^{-1}\mathfrak{S}_\gamma}\prod_{k=1}^{r}\prod_{l=k+1}^{r}e^{-c\,b_{kl}^{\lambda}(\tau)}\right)\left(\prod_{s=1}^{r}\prod_{j=1}^{\lambda_s-1}\sum_{k=0}^{j}e^{-kc}\right). \tag{16}$$

Proof Using $(\mathfrak{S}_\gamma\pi)\tau = \mathfrak{S}_\gamma(\pi\tau)$, Corollary 1, and the fact that $\tau\in\mathfrak{S}_\gamma$ iff $\tau^{-1}\in\mathfrak{S}_\gamma$, we have

$$\begin{aligned} \sum_{\sigma\in\mathfrak{S}_\lambda\pi_1}\sum_{\tau\in\mathfrak{S}_\gamma\pi_2}e^{-cd(\sigma,\tau)} &= \sum_{\sigma\in\mathfrak{S}_\lambda}\sum_{\tau\in\mathfrak{S}_\gamma}e^{-cd(\sigma\pi_1,\tau\pi_2)} = \sum_{\sigma\in\mathfrak{S}_\lambda}\sum_{\tau\in\mathfrak{S}_\gamma}e^{-cd(\sigma\pi_1\pi_2^{-1}\tau^{-1},e)} \\ &= \sum_{\tau\in\mathfrak{S}_\gamma}\sum_{\sigma\in\mathfrak{S}_\lambda}e^{-c\,i(\sigma\pi_1\pi_2^{-1}\tau^{-1})} = \sum_{\tau\in\mathfrak{S}_\gamma}\sum_{\sigma\in\mathfrak{S}_\lambda}e^{-c\,i(\sigma(\pi_1\pi_2^{-1})\tau)} \\ &= \sum_{\tau\in\mathfrak{S}_\gamma}e^{-c\sum_{k=1}^{r}\sum_{l=k+1}^{r}b_{kl}^{\lambda}(\pi_1\pi_2^{-1}\tau)}\prod_{s=1}^{r}\prod_{j=1}^{\lambda_s-1}\sum_{k=0}^{j}e^{-kc}. \end{aligned}$$

■

The complexity of computing (16), (12), (13) for some popular partial ranking types appears in Table 1. The independence of these complexity terms from n enables the practical use of estimators (12), (13) in large n situations. Some of the details concerning this complexity analysis and algorithmic implementation may be found in Appendix A.

8. Statistical Properties of the Estimator

After studying the computational feasibility of the non-parametric estimator $\hat{p}$ in the previous section, we now turn to examine its statistical properties. In particular we examine its bias and variance,

$\lambda \diagdown \gamma$	$(1,n-1)$	$(1,\cdots,1,n-t)$	$(t,n-t)$
$(1,n-1)$	$O(1)$	$O(1)$	$O(1)$
$(1,\cdots,1,n-k)$	$O(k)$	$O(k+t)$	$O(k+t)$
$(k,n-k)$	$O(k)$	$O(k+t)$	$O(k+t)$

Table 1: Computational complexity for computing Equation (13) for each training example. The independence of the complexity terms from n enables the practical use of the estimators (12),(13) in $k \ll n$ situations.

show consistency for large n, and examine the statistical effect of using partially ranked or censored data in the estimation process. Due to the discreteness of the probability space we replace traditional Taylor series expansion with a bound based on the Lipschitz continuity of p. The Lipschitz continuity assumption is crucial since without such an assumption on the regularity of p, kernel based smoothing or other neighborhood operations make little sense.

Proposition 7 *Let $\pi_1,\ldots,\pi_m \in \mathfrak{S}_n$ be sampled iid from a Lipschitz continuous p, that is, $|p(\pi)-p(\tau)| \leq M\,d(\pi,\tau)$, $\forall \pi,\tau$. The following bounds with respect to $\hat{p}$ in* (11) *hold.*

$$|\mathit{bias}\,(\hat{p}(\pi))| \leq -M\frac{\psi'(c)}{\psi(c)},$$
$$\mathit{Var}\,(\hat{p}(\pi)) \leq \frac{p(\pi)}{m}\frac{\psi(2c)}{\psi^2(c)} - \frac{M}{m}\frac{\psi'(2c)}{\psi^2(c)}.$$

Proof Key properties in the following manipulations are the closed form expression of $\psi(c)$ in (6) and its independence from the location parameter of the Mallows kernel.

$$\begin{aligned}
|\mathsf{bias}\,(\hat{p}(\pi))| &= \left|\mathsf{E}_{p(\pi_1)}\left(\psi^{-1}(c)e^{-cd(\pi,\pi_1)}\right) - p(\pi)\right| \\
&\leq \psi^{-1}(c)\sum_{\pi_1\in\mathfrak{S}_n}|p(\pi_1)-p(\pi)|e^{-cd(\pi,\pi_1)} \\
&\leq \psi^{-1}(c)\sum_{\pi_1\in\mathfrak{S}_n} Md(\pi,\pi_1)e^{-cd(\pi,\pi_1)} = -M\frac{\psi'(c)}{\psi(c)}.
\end{aligned}$$

$$\begin{aligned}
\psi^2(c)\,m\,\mathsf{Var}\,(\hat{p}(\pi)) &= \mathsf{Var}_{\,p(\pi_1)}e^{-c\,d(\pi,\pi_1)} \leq \mathsf{E}_{\,p(\pi_1)}e^{-2cd(\pi,\pi_1)} \\
&= \sum_{\pi_1\in\mathfrak{S}_n} p(\pi_1)e^{-2c\,d(\pi,\pi_1)} \\
&\leq \sum_{\pi_1\in\mathfrak{S}_n} (p(\pi)+M\,d(\pi,\pi_1))e^{-2cd(\pi,\pi_1)} \\
&= p(\pi)\psi(2c) - M\psi'(2c).
\end{aligned}$$

■

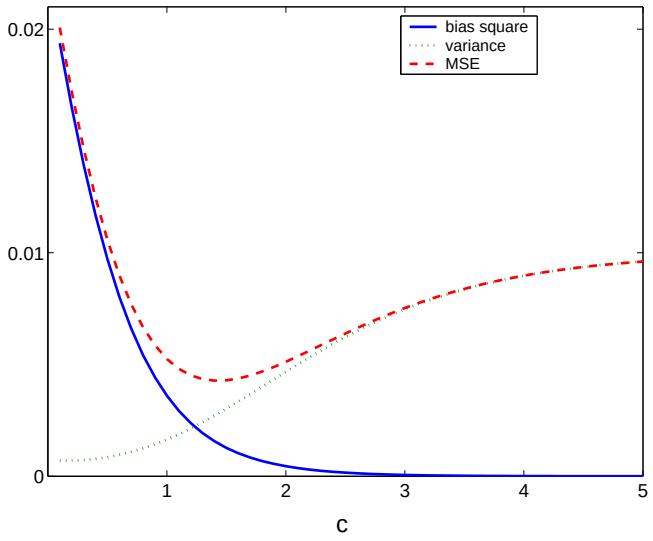

Figure 4: Upper bounds on squared bias, variance and MSE as functions of c: $M = 0.05$, $p(\pi) = 0.2$, $n = 4$, $m = 20$.

The upper bounds in Proposition 7 are illustrated as functions of c in Figure 4. These expressions may be written in a closed form using the formulas for $\psi(2c)/\psi^2(c)$ and $\psi'(c)/\psi(c)$ derived in the proof of Proposition 8 below.

Proposition 8 *Under the same conditions as Proposition 7 and assuming the asymptotics*

$$c, m, n \to \infty, \quad n = o(\exp(c)), \quad n = o(\sqrt{m})$$

the estimator $\hat{p}$ in (11) *is pointwise consistent.*

Proof We first derive closed form expressions for $\psi'(c)/\psi(c)$ and $\psi(2c)/\psi^2(c)$ and then proceed to demonstrate the convergence to 0 of the bias and variance bounds obtained in Proposition 7.

Using the result $\psi(c) = \prod_{j=1}^{n} \frac{1-e^{-jc}}{1-e^{-c}}$ shown in (6), we have

$$\frac{\psi'(c)}{\psi(c)} = (\log \psi(c))' = \sum_{j=1}^{n} \frac{j e^{-jc}}{1-e^{-jc}} - \frac{n e^{-c}}{1-e^{-c}}, \tag{17}$$

$$\begin{aligned} \frac{\psi(2c)}{\psi^2(c)} &= \prod_{j=1}^{n} \frac{1-e^{-2jc}}{1-e^{-2c}} \frac{(1-e^{-c})^2}{(1-e^{-jc})^2} = \prod_{j=1}^{n} \frac{1+e^{-jc}}{1+e^{-c}} \frac{1-e^{-jc}}{1-e^{-c}} \frac{(1-e^{-c})^2}{(1-e^{-jc})^2} \\ &= \prod_{j=1}^{n} \frac{1+e^{-jc}}{1-e^{-jc}} \frac{1-e^{-c}}{1+e^{-c}}. \end{aligned} \tag{18}$$

The term $-\psi'(c)/\psi(c)$ is the expected distance under the Mallows model

$$-\psi'(c)/\psi(c) = \sum_{\sigma \in \mathfrak{S}_n} d(\pi, \sigma) \psi^{-1}(c) \exp(-c\, d(\pi, \sigma))$$

and therefore is bounded by $\max_{\pi,\sigma} d(\pi,\sigma) \leq n^2$. The term $\psi(2c)/\psi^2(c)$ is bounded since it may be written as a product $\prod_{j=1}^{n} R_j(c)$, with $R_j(c) = \frac{1+e^{-jc}}{1-e^{-jc}} / \frac{1+e^{-c}}{1-e^{-c}} \leq 1$ for all $c \in \mathbb{R}_+$ and $j \geq 1$ since the function $\frac{1+\varepsilon}{1-\varepsilon}$ increases with $\varepsilon > 0$.

Based on Proposition 7 and Equations (17)-(18)

$$|\mathrm{bias}\,(\hat{p}(\pi))| \leq M\frac{ne^{-c}}{1-e^{-c}} - M\sum_{j=1}^{n}\frac{je^{-jc}}{1-e^{-jc}} \leq M\frac{ne^{-c}}{1-e^{-c}}$$

$$\mathrm{Var}\,(\hat{p}(\pi)) \leq \frac{p(\pi)}{m}\frac{\psi(2c)}{\psi^2(c)} - \frac{M}{m}\frac{\psi'(2c)}{\psi(2c)}\frac{\psi(2c)}{\psi^2(c)}.$$

The bias converges to 0 as $n\exp(-c) \to 0$ or alternatively, $c \to \infty, n = o(\exp(c))$. Since $\psi(2c)/\psi^2(c)$ is bounded and $-\psi'(2c)/\psi(2c) \leq n^2$ the variance converges to 0 as well if $m \to \infty$ and $n^2/m \to 0$. ■

Intuitively, the inverse scale parameter c has to go to ∞ in order for the bias to converge to 0 (similar to the requirement $h \to 0$ for the bandwidth parameter h in kernel smoothing). The number of samples m has to go to ∞ in order for the variance to go to 0. Allowing $n \to \infty$ enables us to study the behavior of $\hat{p}$ in situations containing a large number of items. The proposition above (with a slightly modified proof) also holds for fixed n.

The assumption above of Lipschitz continuity with respect to d is a very weak assumption since the distance d tends to grow as $n \to \infty$. In particular d takes values in $[0, n(n-1)/2]$ making the Lipschitz continuity assumption weaker and weaker as $n \to \infty$. A stronger assumption of Lipschitz continuity with respect to the normalized d_n (5)

$$|p(\pi) - p(\tau)| \leq M\, d_n(\pi,\tau), \quad \forall \pi,\tau$$

results in a similar conclusion to Proposition 8 asserting pointwise consistency of $\hat{p}$ under weaker asymptotic requirements.

For large n, it is often the case that partial, rather than full, rankings are available for estimating $\hat{p}$. Partially ranked data is easier for rankers to express than a lengthy list corresponding to a precise permutation. Furthermore, in many cases, rankers can make some partial ranking assertions with certainty but do not have a clear opinion on other preferences. Using the censored data interpretation of partially ranked data enables efficient use of partially ranked data of multiple types in the estimation process (12).

Statistically, expressing partially ranked data as censored data has the effect of increased smoothing and therefore it reduces the variance while increasing the bias. The following proposition quantifies this effect in terms of the bias and variance of $\hat{p}$. A consequence of this proposition which is also illustrated in Section 9 experimentally is that even if the fully ranked data is somehow available, estimating $\hat{p}$ based on the partial rankings obtained by censoring it tends to increase the estimation accuracy.

Proposition 9 *Assuming the same conditions as in Proposition 7, the bias and variance of the censored data or partial ranking estimator* (12) *for* $\gamma_1 = \ldots = \gamma_m = \gamma$ *satisfy*

$$|bias\,(\hat{p}(\pi))| \leq -M\frac{\psi'(c)}{\psi(c)} + M\frac{sp(\mathfrak{S}_\gamma)}{|\mathfrak{S}_\gamma|},$$
$$Var\,(\hat{p}(\pi)) \leq \frac{p(\pi)}{m}\frac{1}{|\mathfrak{S}_\gamma|} + \frac{M}{m}\frac{sp(\mathfrak{S}_n)}{|\mathfrak{S}_\gamma|^2} \tag{19}$$

where $sp(U) \stackrel{\text{def}}{=} \max_{x\in U}\sum_{y\in U} d(x,y)$.

The choice of using $\gamma_1 = \ldots = \gamma_m = \gamma$ above was made for simplicity. Similar results apply for more heterogenous partially ranked data.

Proof

$$\begin{aligned}
|\mathrm{bias}\,(\hat{p}(\pi))| &= \left|\psi^{-1}(c)|\mathfrak{S}_\gamma|^{-1}\mathsf{E}_{p(\pi_1)}\sum_{\tau\in\mathfrak{S}_\gamma\pi_1} e^{-cd(\pi,\tau)} - p(\pi)\right| \\
&\stackrel{a}{\leq} \psi^{-1}(c)|\mathfrak{S}_\gamma|^{-1}\sum_{\pi_1\in\mathfrak{S}_n}\sum_{\tau\in\mathfrak{S}_\gamma\pi_1}|p(\pi_1)-p(\pi)|e^{-cd(\pi,\tau)} \\
&\leq M\psi^{-1}(c)|\mathfrak{S}_\gamma|^{-1}\sum_{\pi_1\in\mathfrak{S}_n}\sum_{\tau\in\mathfrak{S}_\gamma\pi_1} d(\pi,\pi_1)e^{-cd(\pi,\tau)} \\
&\leq M\psi^{-1}(c)|\mathfrak{S}_\gamma|^{-1}\sum_{\pi_1\in\mathfrak{S}_n}\sum_{\tau\in\mathfrak{S}_\gamma\pi_1} (d(\pi,\tau)+d(\tau,\pi_1))e^{-cd(\pi,\tau)} \\
&\stackrel{b}{=} -M\frac{\psi'(c)}{\psi(c)} + \frac{M}{\psi(c)|\mathfrak{S}_\gamma|}\sum_{\pi_1\in\mathfrak{S}_n}\sum_{\tau\in\mathfrak{S}_\gamma\pi_1} d(\tau,\pi_1)e^{-cd(\pi,\tau)}
\end{aligned}$$

where *a* and *b* follow from the fact that

$$\sum_{\pi_1\in\mathfrak{S}_n}\sum_{\tau\in\mathfrak{S}_\gamma\pi_1} e^{-cd(\pi,\tau)} = |\mathfrak{S}_\gamma|\sum_{\tau\in\mathfrak{S}_n} e^{-cd(\pi,\tau)} = |\mathfrak{S}_\gamma|\psi(c).$$

The inner summation depends on π_1 only through the coset $\mathfrak{S}_\gamma\pi_1$ it resides in. To simplify the expression, we separate the single outer summation to summations of π_1 over the distinct cosets C_j. Since the number of distinct $\mathfrak{S}_\gamma$ cosets in $\mathfrak{S}_n$ is the index $[\mathfrak{S}_n : \mathfrak{S}_\gamma] = |\mathfrak{S}_n|/|\mathfrak{S}_\gamma|$, we have

$$\begin{aligned}
|\mathrm{bias}\,(\hat{p}(\pi))| &\leq -M\frac{\psi'(c)}{\psi(c)} + \frac{M}{\psi(c)|\mathfrak{S}_\gamma|}\sum_{j=1}^{[\mathfrak{S}_n:\mathfrak{S}_\gamma]}\sum_{\tau\in C_j}\left(\sum_{\pi_1\in C_j} d(\tau,\pi_1)\right)e^{-cd(\pi,\tau)} \\
&\leq -M\frac{\psi'(c)}{\psi(c)} + \frac{M\,\mathrm{sp}(\mathfrak{S}_\gamma)}{\psi(c)|\mathfrak{S}_\gamma|}\sum_{j=1}^{[\mathfrak{S}_n:\mathfrak{S}_\gamma]}\sum_{\tau\in C_j} e^{-cd(\pi,\tau)} \\
&= -M\frac{\psi'(c)}{\psi(c)} + M\frac{\mathrm{sp}(\mathfrak{S}_\gamma)}{|\mathfrak{S}_\gamma|}
\end{aligned}$$

using the fact that the spread is the same for all cosets of the same type $\mathrm{sp}(\mathfrak{S}_\gamma\pi) = \mathrm{sp}(\mathfrak{S}_\gamma)$.

$$
\begin{aligned}
m\psi^2(c)\,|\mathfrak{S}_\gamma|^2\,\mathsf{Var}\,(\hat{p}(\pi)) &= \mathsf{Var}_{\,p(\pi_1)} \sum_{\tau\in\mathfrak{S}_\gamma\pi_1} e^{-cd(\pi,\tau)} \le \mathsf{E}_{\,p(\pi_1)} \left(\sum_{\tau\in\mathfrak{S}_\gamma\pi_1} e^{-cd(\pi,\tau)} \right)^2 \\
&\le \sum_{\pi_1\in\mathfrak{S}_n} (p(\pi) + Md(\pi,\pi_1)) \left(\sum_{\tau\in\mathfrak{S}_\gamma\pi_1} e^{-cd(\pi,\tau)} \right)^2 \\
&= \sum_{j=1}^{[\mathfrak{S}_n:\mathfrak{S}_\gamma]} \left(\sum_{\tau\in C_j} e^{-cd(\pi,\tau)} \right)^2 \left(p(\pi)|\mathfrak{S}_\gamma| + M \sum_{\sigma\in C_j} d(\pi,\sigma) \right) \\
&\le p(\pi)|\mathfrak{S}_\gamma|\psi^2(c) + M\mathrm{sp}(\mathfrak{S}_n)\psi^2(c).
\end{aligned}
$$

In the last inequality we used the Cauchy-Schwartz inequality $\langle u,v\rangle \le \|u\|_2\|v\|_2 \le \|u\|_1\|v\|_1$ to obtain

$$
\sum_{j=1}^{[\mathfrak{S}_n:\mathfrak{S}_\gamma]} \left(\sum_{\tau\in C_j} e^{-cd(\pi,\tau)} \right)^2 \le \left(\sum_{j=1}^{[\mathfrak{S}_n:\mathfrak{S}_\gamma]} \sum_{\tau\in C_j} e^{-cd(\pi,\tau)} \right)^2 = \psi^2(c).
$$

■

Contrasting the expressions in Proposition 9 with those in Proposition 7 indicates that reverting to partial rankings tends to increase the bias but reduce the variance. Intuitively, the bias increases since we no longer have enough data, in general, to precisely estimate the permutation model p. The variance (19), on the other hand, experiences a substantial reduction as compared to the fully ranked case. Figure 5 displays the behavior of the quantities $\frac{\mathrm{sp}(\mathfrak{S}_\gamma)}{|\mathfrak{S}_\gamma|}$ and $\frac{\mathrm{sp}(\mathfrak{S}_n)}{|\mathfrak{S}_\gamma|^2}$. The first quantity $\frac{\mathrm{sp}(\mathfrak{S}_\gamma)}{|\mathfrak{S}_\gamma|}$, which bounds the bias, increases as the composition γ represents a lower degree of specificity. On the other hand, the second quantity $\frac{\mathrm{sp}(\mathfrak{S}_n)}{|\mathfrak{S}_\gamma|^2}$ which bounds the variance decreases as the composition γ represents less specificity.

The precise changes in the bias and variance that occur due to using partial rankings depend on γ, n, m, c, M. However, generally speaking, the variance reduction becomes more pronounced as n and $|\mathfrak{S}_\gamma|$ grow. Indeed, in the common case described earlier where the number of items n is large, switching to partially ranked data can dramatically improve the estimation accuracy. This observation, which is illustrated in Section 9 using experiments on real world data, becomes increasingly important as n increases. It is remarkable that this statistical motivation to use partial rather than full rankings is aligned with the data availability and ease of use as well as with the computational efficiency demonstrated in the previous section.

9. Applications and Experiments

The estimator $\hat{p}$ defined in (11), (12) and its lattice version $\hat{g}$ defined in (13) can be used in a number of data analysis tasks. We briefly outline some of these tasks below and then proceed to describe some experimental results.

Visual or computational exploration of the model probabilities $\{\hat{p}(\pi) : \pi \in \mathfrak{S}_n\}$ can be a useful exploratory data analysis tool. Such exploration can be done by visualizing the values $\{\hat{p}(\pi) : \pi \in \mathfrak{S}_n\}$ for small n using the techniques developed in Thompson (1993). For medium and large n

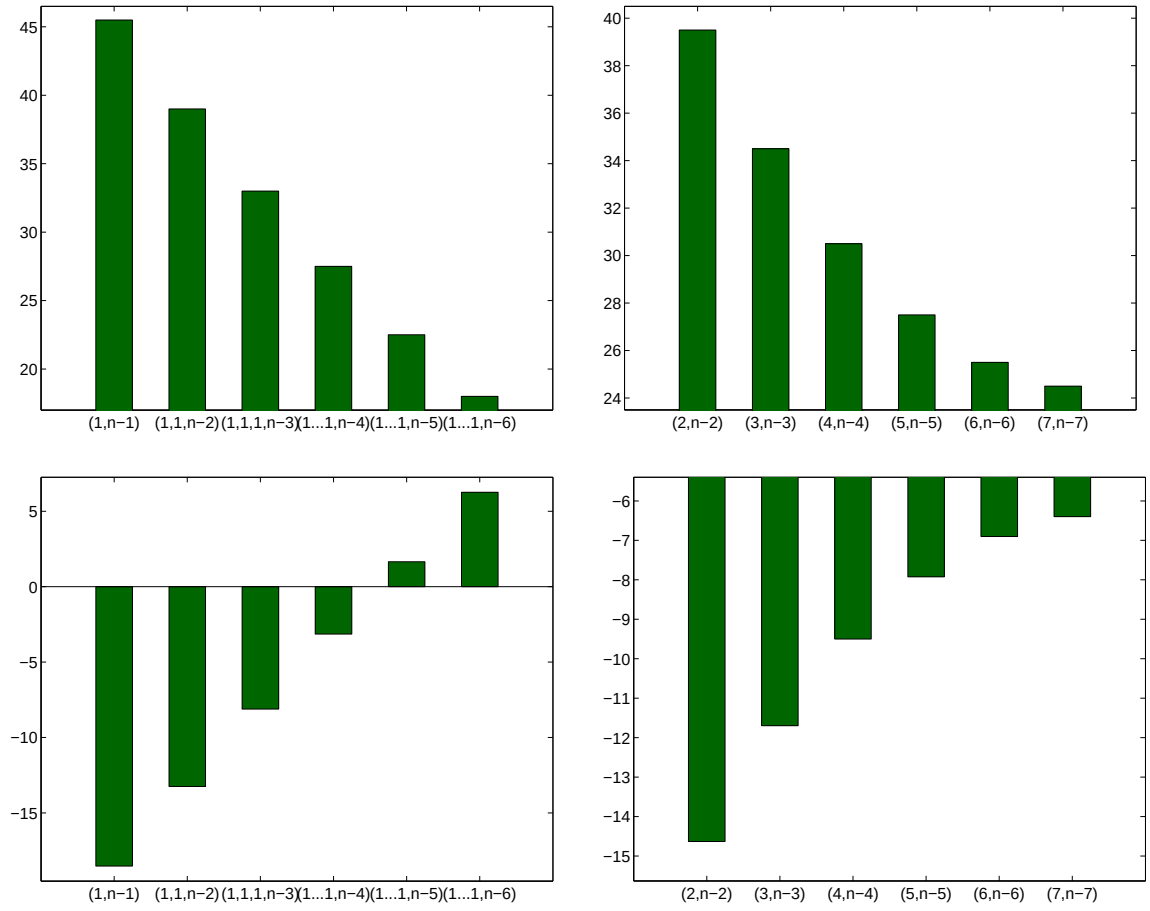

Figure 5: Values of $\frac{\mathrm{sp}(\mathfrak{S}_\gamma)}{|\mathfrak{S}_\gamma|}$(top row), and $\log \frac{\mathrm{sp}(\mathfrak{S}_n)}{|\mathfrak{S}_\gamma|^2}$ (bottom row) for $n = 15$ and various partial ranking types. Note that $\frac{\mathrm{sp}(\mathfrak{S}_\gamma)}{|\mathfrak{S}_\gamma|}$ (which serves as a bound for the bias) decreases and $\frac{\mathrm{sp}(\mathfrak{S}_n)}{|\mathfrak{S}_\gamma|^2}$ (which serves as a bound for the variance) increases for decreasing $|\mathfrak{S}_\gamma|$.

similar visualization techniques can be used to explore the values of the lattice version $\hat{g}$ restricted to certain subset $H \subset \tilde{\mathfrak{W}}_n$ of the ranking lattice. Since the number of distinct γ-cosets $|\mathfrak{S}_n|/|\mathfrak{S}_\gamma|$ may be much smaller than $|\mathfrak{S}_n|$, visualizing $\{\hat{g}(A) : A \in H\}$ can be more effective than visualizing $\{\hat{p}(\pi) : \pi \in \mathfrak{S}_n\}$. Other explorations such as identifying the local modes of $\hat{p}$ and $\hat{g}$ may be automated and computed without human intervention.

In some cases, the main objective of inference is a conditional version of $\hat{p}$ such as $\hat{p}(\pi \in A | \pi \in B)$, $A, B \subset \mathfrak{S}_n$. A popular example is collaborative filtering which is the task of recommending items to a user based on partial preference information that is output by that user (Resnick et al., 1994). In this case, $\hat{p}$ is estimated based on a large data set of partial preferences provided by many users. Given a particular partial ranking $\mathfrak{S}_\gamma\pi$ output by a certain user we can predict its most likely refinement $\arg\max_{\mathfrak{S}_\lambda\sigma} \hat{p}(\mathfrak{S}_\lambda\sigma | \mathfrak{S}_\gamma\pi)$. This task is central to many recommendation systems and

has recently gained popularity in the machine learning research community due to its commercial applications.

Statistics such as expectations and variances can be useful as summaries in situations where the entire distribution is not necessary. For example, summaries such as the expectation and variance of an item's rank $\mathsf{E}_{\hat{p}}(\pi(k))$, $\mathsf{Var}_{\hat{p}}(\pi(k))$ or probabilities such as $\hat{p}(\pi(i) > \pi(j))$ may be useful in some cases. On the other hand, in situations where $\hat{p}$ is a complex multimodal distribution, the summaries need to be complemented with a careful examination of $\hat{p}$.

We experimented with three different data sets. The first is the APA data set (Diaconis, 1989) which contains several thousand rankings of 5 APA presidential candidates. The second is the Jester data set containing rankings of 100 jokes by 73,496 users. The third data set is the EachMovie data set containing rankings of 1628 movies by 72,916 users. In our experiments, we trained models based on a randomly sampled training set and evaluated the log-likelihood on a separate held-out testing set. We repeated this procedure 10 times and report the average log-likelihood in order to reduce sampling noise.

Figure 6 displays the test set log-likelihood for the parametric Mallows model (fitted by maximum likelihood) and the non-parametric estimator. The log-likelihood, computed as a function of the train set size, is displayed for several values of c for the non-parametric estimator. In the case of the Mallows model we only display the optimal c. Due to the computational difficulty associated with maximum likelihood for the Mallows model for large n we experimented with rankings over a small number of items. The three panels of the figure display the log-likelihood with respect to the APA data with $n = 5$ (top), the Jester data restricted to the $n = 5$ most frequently rated jokes (middle), and the EachMovie data restricted to the $n = 4$ most frequently rated movies. In all three cases, the non-parametric estimator performed better than the parametric Mallows model given sufficient training examples. As c increases, the non-parametric model tends to perform better for large data sets and worse for small data sets, reflecting the non-parametric consistency as $m, c \to \infty$.

The increased flexibility of the non-parametric model illustrated in Figure 6 can be visualized further by comparing the probabilities assigned by the Mallows model and the non-parametric model. We display these probabilities in the case of $n = 4$ (movies no. 357, 1356, 440, 25 from the EachMovie data) by scaling appropriately the vertices of the permutation polytope. The vertices of the permutation polytope, displayed in Figure 7, correspond to $\mathfrak{S}_4$ and its edges correspond to pairs of permutations with Kendall's tau distance 1. In fact, Kendall's tau distance $d(\pi, \sigma)$ corresponds to the length of the shortest path connecting the two vertices representing π and σ. As a result, the 3D embedding of the permutation polytope effectively visualizes the discrete metric space $(\mathfrak{S}_4, d)$. In the figure, the radiuses of the vertices were scaled proportionally to $(\hat{p}(\pi))^{5/7}$ where $\hat{p}(\pi)$ are the probabilities estimated by maximum likelihood Mallows model (left) and the non-parametric model (right). The scaling exponent of $5/7$ was chosen in agreement with Steven's law (Cleveland, 1985) for effective visualization. Figure 7 shows that the probabilities assigned by the Mallows model form a diffuse unimodal function centered at $2|1|3|4$. The non-parametric estimator, on the other hand, discovers the true global mode $2|3|1|4$ and an additional local mode at $4|1|2|3$ both of which were undiscovered by the Mallows model due to its unimodality property.

Figure 8 demonstrates non-parametric modeling of partial rankings for $n = 100$ (the Mallows model maximum likelihood estimator cannot be computed for such n). We used 10043 rankings from the Jester data set which contain users ranking all $n = 100$ jokes. As before, the figures display the test-set log-likelihood as a function of the train set size. Due to the large n, we measured the test

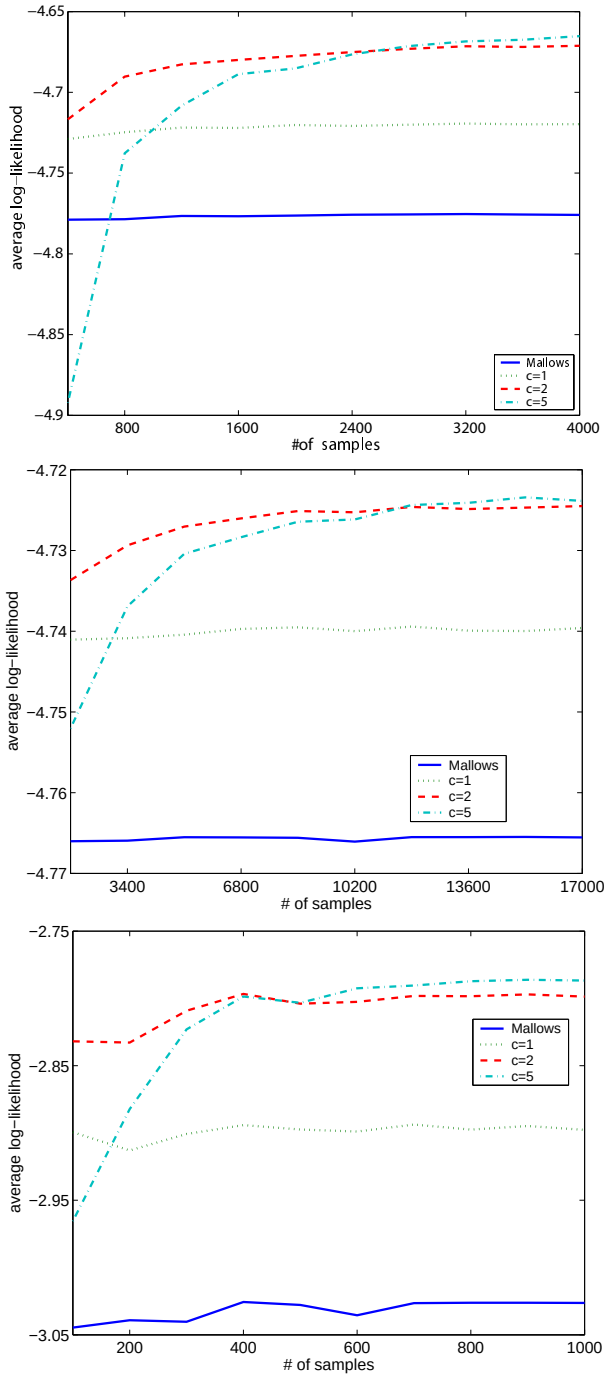

Figure 6: Average test log-likelihood as a function of the train set size: the maximum likelihood Mallows model vs. the non-parametric estimator for (a) APA data $n = 5$, (b) $n = 5$ most frequently rated Jester jokes, (c) $n = 4$ most frequently rated movies from EachMovie data. In general, the non-parametric model provides a better fit than the Mallows model. The non-parametric consistency is illustrated in the case of $c, m \to \infty$.

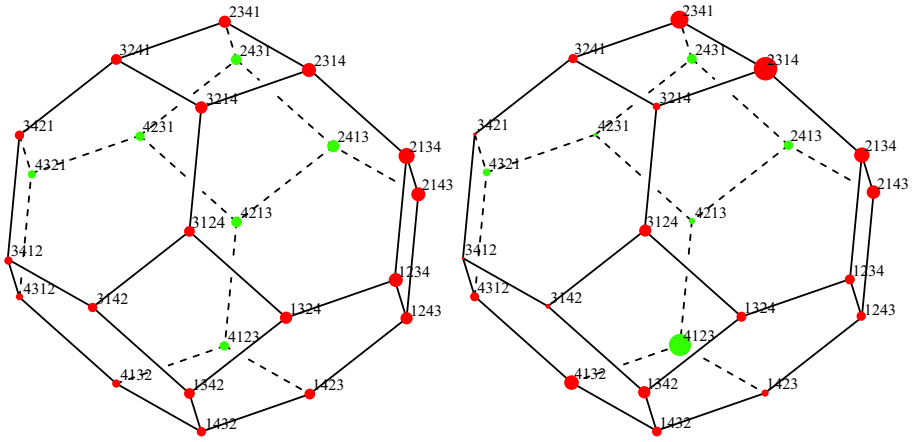

Figure 7: Visualizing estimated probabilities for EachMovie data by permutation polytopes: Mallows model (left) and non-parametric model for $c = 2$ (right). The Mallows model locates a single mode at $2|1|3|4$ while the non-parametric estimator locates the global mode at $2|3|1|4$ and a second local mode at $4|1|2|3$.

set log-likelihood with respect to the lattice version $\hat{g}(\mathfrak{S}_\gamma \pi)$ of the non-parametric estimator $\hat{p}$ for partial ranking $\gamma = (5, n-5)$ (top) and $\gamma = (1,1,1,n-3)$ (bottom).

The different lines in Figure 8 correspond to the performance of $\hat{p}$ obtained by censoring the training data in different ways. We compared $\hat{p}$ for the following censored data: full ranking (no censoring), $\gamma = (1,\ldots,1,n-k)$ for $k = 1,2,3,5$ and $\gamma = (k, n-k)$. The value of k in the censoring corresponding to $\gamma = (k, n-k)$ was chosen based on thresholding the scores output by the users. In particular, $(k(s), n-k(s))$ corresponds to k being the number of jokes receiving a score of s or higher (the users provided scores in the range $[-10, 10]$). The figure illustrates the statistical benefit of estimating $\hat{p}$ based on partial rather than full rankings. The variance reduction by $(k, n-k)$ partial rankings clearly outweighs the bias increase.

10. Discussion

As the number of items n increases, the space $\mathfrak{S}_n$ grows exponentially making discrete search methods such as maximum likelihood for the Mallows model difficult to compute. Similarly, it is typically the case for large n that both the data available for estimating $\hat{p}$ and the use of $\hat{p}$ will be restricted to partial rankings or cosets of the symmetric group.

Attempts to define a probabilistic model directly on multiple types of partial rankings $H \subset \tilde{\mathfrak{W}}_n$ face a challenging problem of preventing probabilistic contradictions. A simple solution is to define the partial ranking model $\hat{g}$ in terms of a permutation model $\hat{p}$ through the mechanism of Möbius inversion and censored data interpretation. However, doing so raises computational concerns that often severely limit the practical use of such models for large n.

In this paper, we present a non-parametric kernel smoothing technique that uses the Mallows model as a smoothing kernel on permutations. Using combinatorial properties of inversions and of the symmetric group we simplify the computational difficulties and exhibit its practical use inde-

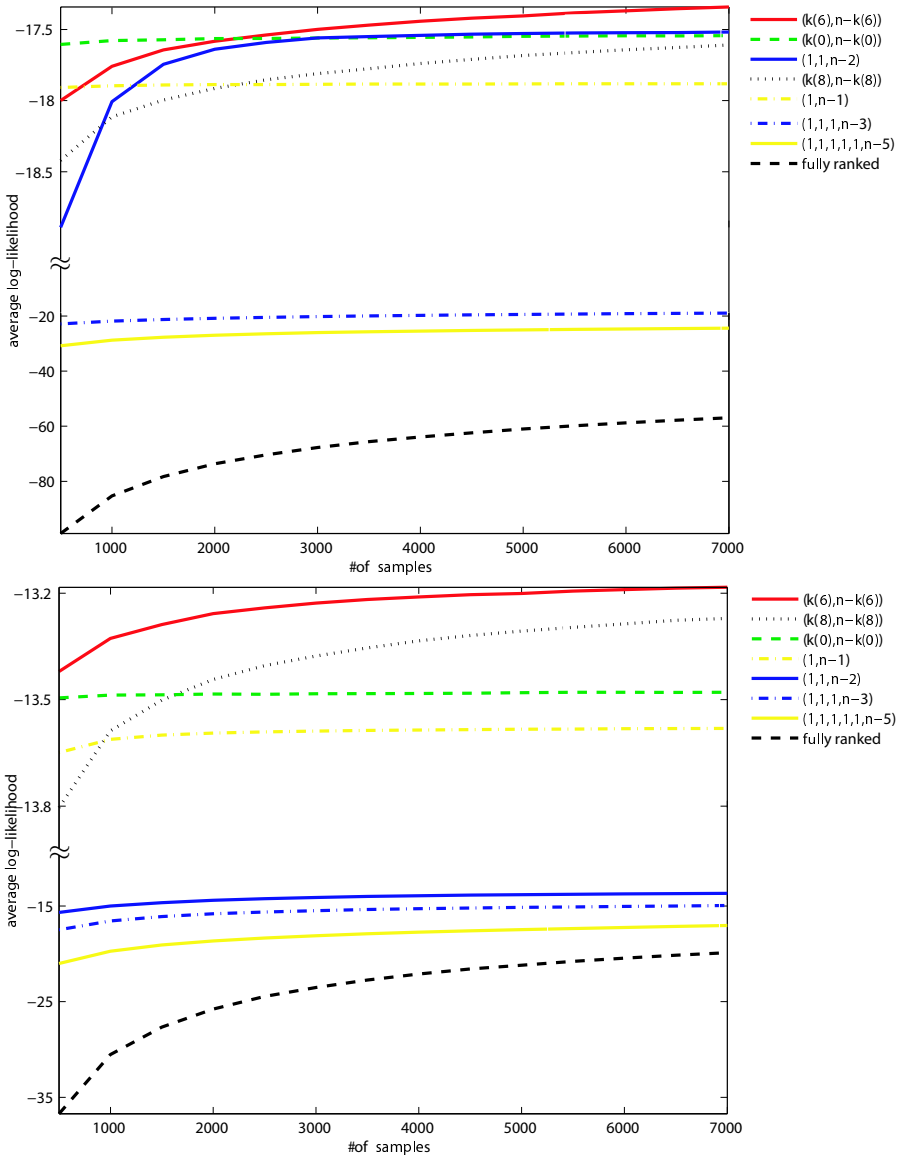

Figure 8: Test-set loglikelihood for $\hat{g}(\mathfrak{S}_\gamma\pi)$ with $\gamma=(5,n-5)$ (top) and $\gamma=(1,1,1,n-3)$ (bottom) as a function of train set size (Jester data set with $n=100$). The different lines correspond to obtaining $\hat{p}$ based on different censoring strategies of the fully ranked training data (see description in text). The legend entries are sorted in roughly the same order as the lines in the figures for increased visibility.

pendently of the number of items n. Theoretical and experimental examinations demonstrate the role of the inverse scale parameter c in the bias-variance tradeoff. We also examine the effect of using partial, rather than full, rankings on the bias and variance of the estimator. This effect plays a similar role to increased kernel smoothing and often leads to increased estimation accuracy.

Appendix A. Complexity Issues

Table 1 lists the computational complexity results for computing (13) for some popular partial ranking types γ and λ. The arguments or proofs for these expressions are rather involved and contain some details. We include in this appendix the details corresponding to the case of $\lambda = (k, n-k)$ and $\gamma = (t, n-t)$. The other cases in Table 1 follow similarly, but with some differences.

Proposition 10 *The complexity for computing*

$$\psi^{-1}(c)\,|\mathfrak{S}_\gamma|^{-1} \sum_{\sigma \in \mathfrak{S}_\lambda \pi_1} \sum_{\tau \in \mathfrak{S}_\gamma \pi_2} e^{-cd(\sigma,\tau)}$$

for $\lambda = (k, n-k)$ and $\gamma = (t, n-t)$ is $O(k+t)$.

Proof We first generalize the definition of cross compartment inversions $b^{\gamma}_{kl}(\tau)$ in Proposition 4 by defining

$$b_{XY}(\tau) = |\{(u,v) : u < v, \tau(v) \in X, \tau(u) \in Y\}|$$

where X and Y are arbitrary disjoint sets. If $X = \left\{ \sum_{j=1}^{k-1} \gamma_j + 1, \ldots, \sum_{j=1}^{k} \gamma_j \right\}$ and $Y = \left\{ \sum_{j=1}^{l-1} \gamma_j + 1, \ldots, \sum_{j=1}^{l} \gamma_j \right\}$, we have $b_{XY}(\tau) = b^{\gamma}_{kl}(\tau)$. If $x < y, \forall x \in X$ and $y \in Y$, $b_{XY}(\tau)$ counts a subset of inversion pairs of τ. However, in its most general form, $b_{XY}(\tau)$ may include non-inversion pairs if some numbers in X are greater than some numbers in Y.

We use the following definitions in our proof

$$\begin{aligned}
A &= \{1, \cdots, k\} \cap \{\pi_1 \pi_2^{-1}(1), \cdots, \pi_1 \pi_2^{-1}(t)\}, \\
\bar{A} &= \{1, \cdots, k\} \setminus A, \\
B &= \{k+1, \cdots, n\} \cap \{\pi_1 \pi_2^{-1}(1), \cdots, \pi_1 \pi_2^{-1}(t)\}, \\
\bar{B} &= \{k+1, \cdots, n\} \setminus B.
\end{aligned}$$

Note $A, \bar{A}, B, \bar{B}$ constitute a partition of $\{1, \ldots, n\}$ and satisfy

$$\begin{aligned}
A \cup \bar{A} &= \{1, \ldots, k\}, \\
B \cup \bar{B} &= \{k+1, \ldots, n\}, \\
A \cup B &= \{\pi_1 \pi_2^{-1}(1), \cdots, \pi_1 \pi_2^{-1}(t)\}, \\
\bar{A} \cup \bar{B} &= \{1, \cdots, n\} \setminus \{\pi_1 \pi_2^{-1}(1), \cdots, \pi_1 \pi_2^{-1}(t)\}.
\end{aligned}$$

Since $\lambda = (k, n-k)$, we have $b^{\lambda}_{12}(\tau) = b_{AB}(\tau) + b_{A\bar{B}}(\tau) + b_{\bar{A}B}(\tau) + b_{\bar{A}\bar{B}}(\tau)$, and the expression in the first parenthesis of Equation 16 is simplified to be

$$\begin{aligned}
\sum_{\tau\in\pi_1\pi_2^{-1}\mathfrak{S}_\gamma}\prod_{k=1}^{r}\prod_{l=k+1}^{r} e^{-cb^{\lambda}_{kl}(\tau)} &= \sum_{\tau\in\pi_1\pi_2^{-1}\mathfrak{S}_\gamma} e^{-cb^{\lambda}_{12}(\tau)} \\
&= \sum_{\tau\in\pi_1\pi_2^{-1}\mathfrak{S}_\gamma} e^{-c(b_{AB}(\tau)+b_{A\bar{B}}(\tau)+b_{\bar{A}B}(\tau)+b_{\bar{A}\bar{B}}(\tau))} \\
&= \sum_{\tau\in\pi_1\pi_2^{-1}\mathfrak{S}_\gamma} e^{-c(b_{AB}(\tau)+0+|\bar{A}||B|+b_{\bar{A}\bar{B}}(\tau))} \\
&= e^{-c|\bar{A}||B|}\sum_{\tau\in\pi_1\pi_2^{-1}\mathfrak{S}_\gamma} e^{-c(b_{AB}(\tau)+b_{\bar{A}\bar{B}}(\tau))} \\
&= e^{-c|\bar{A}||B|}\sum_{\tau\in\mathfrak{S}_t} e^{-cb^{\gamma_1}_{12}(\tau)}\sum_{\tau\in\mathfrak{S}_{n-t}} e^{-cb^{\gamma_2}_{12}(\tau)}
\end{aligned}$$

where $\gamma_1 = (|A|, t-|A|)$ and $\gamma_2 = (|\bar{A}|, n-t-|\bar{A}|)$. The last equality comes from the fact that $\forall\tau \in \pi_1\pi_2^{-1}\mathfrak{S}_\gamma$

$$\begin{aligned}
&\tau^{-1}(i) \in \{1,\ldots,t\} && \text{if} \quad i \in A\cup B, \\
&\tau^{-1}(i) \in \{t+1,\ldots,n\} && \text{if} \quad i \in \bar{A}\cup\bar{B}
\end{aligned}$$

and the choice of representatives π_1, π_2 of the cosets $\mathfrak{S}_\lambda\pi_1, \mathfrak{S}_\gamma\pi_2$ does not change $|A|, |\bar{A}|, |B|$ or $|\bar{B}|$.

By Proposition 11, we have

$$\begin{aligned}
\sum_{\tau\in\mathfrak{S}_t} e^{-cb^{\gamma_1}_{12}(\tau)} &= \frac{|A|!(t-|A|)!\prod_{j=t-|A|+1}^{t}(1-e^{-jc})}{\prod_{j=1}^{|A|}(1-e^{-jc})} \\
\sum_{\tau\in\mathfrak{S}_{n-t}} e^{-cb^{\gamma_2}_{12}(\tau)} &= \frac{|\bar{A}|!(n-t-|\bar{A}|)!\prod_{j=n-t-|\bar{A}|+1}^{n-t}(1-e^{-jc})}{\prod_{j=1}^{|\bar{A}|}(1-e^{-jc})}.
\end{aligned}$$

Substituting the above results into Equation 16, we get

$$\begin{aligned}
&\psi^{-1}(c)|\mathfrak{S}_\gamma|^{-1}\sum_{\sigma\in\mathfrak{S}_\lambda\pi_1}\sum_{\tau\in\mathfrak{S}_\gamma\pi_2} e^{-cd(\sigma,\tau)} \\
&= \frac{\left(e^{-c|\bar{A}||B|}\sum_{\tau\in\mathfrak{S}_t} e^{-cb^{\gamma_1}_{12}(\tau)}\sum_{\tau\in\mathfrak{S}_{n-t}} e^{-cb^{\gamma_2}_{12}(\tau)}\right)\left(\prod_{j=1}^{k}\frac{1-e^{-jc}}{1-e^{-c}}\prod_{j=1}^{n-k}\frac{1-e^{-jc}}{1-e^{-c}}\right)}{\left(\prod_{j=1}^{n}\frac{1-e^{-jc}}{1-e^{-c}}\right)t!(n-t)!} \\
&= \frac{\left(\frac{|A|!(t-|A|)!\prod_{j=t-|A|+1}^{t}(1-e^{-jc})}{\prod_{j=1}^{|A|}(1-e^{-jc})}\right)\left(\frac{|\bar{A}|!(n-t-|\bar{A}|)!\prod_{j=n-t-|\bar{A}|+1}^{n-t}(1-e^{-jc})}{\prod_{j=1}^{|\bar{A}|}(1-e^{-jc})}\right)e^{-c|\bar{A}||B|}\prod_{j=1}^{k}\left(1-e^{-jc}\right)}{t!(n-t)!\prod_{j=n-k+1}^{n}(1-e^{-jc})} \\
&= \frac{|A|!|\bar{A}|!(t-|A|)!e^{-c|\bar{A}||B|}}{t!\prod_{j=n-t-|\bar{A}|+1}^{n-t} j}\left(\frac{\prod_{j=t-|A|+1}^{t}(1-e^{-jc})\prod_{j=n-t-|\bar{A}|+1}^{n-t}(1-e^{-jc})\prod_{j=1}^{k}\left(1-e^{-jc}\right)}{\prod_{j=1}^{|A|}(1-e^{-jc})\prod_{j=1}^{|\bar{A}|}(1-e^{-jc})\prod_{j=n-k+1}^{n}(1-e^{-jc})}\right).
\end{aligned}$$

Note $|A| \le \min(k,t)$, $|\bar{A}| \le k$ and $|B| \le t$, therefore the above expression takes $O(k+t)$ to evaluate. Assuming π_1^{-1} and π_2^{-1} are given, it takes $O(k)$ to get a representative π_1 for the coset $\mathfrak{S}_\lambda\pi_1$, and

$O(t)$ to get the set $\{\pi_1\pi_2^{-1}(1),\cdots,\pi_1\pi_2^{-1}(t)\}$, which completes the proof. ■

Proposition 11 *For $\gamma=(k,n-k)$, let*

$$\mathbf{Q}(k,n) \stackrel{\text{def}}{=} \sum_{\pi\in\mathfrak{S}_n} q^{b_{12}^{\gamma}(\pi)} \tag{20}$$

where $b_{12}^{\gamma}(\pi)$ is defined in Proposition 4, we have

$$\mathbf{Q}(k,n) = k!(n-k)!\frac{\prod_{i=n-k+1}^{n}(1-q^i)}{\prod_{i=1}^{k}(1-q^i)} \quad \forall n\geq k. \tag{21}$$

Proof We first derive an equivalent expression for (20). For fixed π, we sort $\{\pi^{-1}(1),\pi^{-1}(2),\cdots,\pi^{-1}(k)\}$ in ascending order and denote them by $a_1<a_2<\cdots<a_k$. Note that

$$b_{12}^{\gamma}(\pi) = (a_1-1)+(a_2-2)+\cdots+(a_k-k) = \sum_{i=1}^{k}(a_i-i).$$

Due to this observation and since there are $k!(n-k)!$ different permutations for each sequence $(a_1,a_2,\cdots,a_k)$, we have

$$\begin{aligned}\mathbf{Q}(k,n) &= \sum_{\pi\in\mathfrak{S}_n} q^{b_{12}^{\gamma}(\pi)} = k!(n-k)!\sum_{a_k=k}^{n}\sum_{a_{k-1}=k-1}^{a_k-1}\cdots\sum_{a_1=1}^{a_2-1} q^{(a_1+\cdots+a_k-1-\cdots-k)}\\ &= \frac{k!(n-k)!}{q^{\frac{(1+k)k}{2}}}\sum_{a_k=k}^{n} q^{a_k}\sum_{a_{k-1}=k-1}^{a_k-1} q^{a_{k-1}}\cdots\sum_{a_1=1}^{a_2-1} q^{a_1}.\end{aligned} \tag{22}$$

We then prove (21) by mathematical induction on k using the result from (22).

(a) *Base case*: Considering (20) for the case $k=0$ we see that $b_{12}^{(0,n)}(\pi)\equiv 0$ and therefore $\mathbf{Q}(k,n)=n!$ for $n\geq k$. A similar result follows from substituting $k=0$ in the right hand side of (21).

(b) *Inductive step*: Assuming (21) holds $\forall n\geq k$ for some k, we have for $k+1$

$$\begin{aligned}\frac{\mathbf{Q}(k+1,n)}{(k+1)!(n-k-1)!} &\stackrel{a}{=} \frac{1}{q^{\frac{(2+k)(1+k)}{2}}}\sum_{a_{k+1}=k+1}^{n} q^{a_{k+1}}\sum_{a_k=k}^{a_{k+1}-1} q^{a_k}\cdots\sum_{a_1=1}^{a_2-1} q^{a_1}\\ &= \frac{1}{q^{1+k}}\sum_{a_{k+1}=k+1}^{n} q^{a_{k+1}}\left(\frac{1}{q^{\frac{(1+k)k}{2}}}\sum_{a_k=k}^{a_{k+1}-1} q^{a_k}\cdots\sum_{a_1=1}^{a_2-1} q^{a_1}\right)\\ &\stackrel{b}{=} \frac{1}{q^{1+k}}\sum_{a_{k+1}=k+1}^{n} q^{a_{k+1}}\frac{\prod_{i=a_{k+1}-k}^{a_{k+1}-1}(1-q^i)}{\prod_{i=1}^{k}(1-q^i)}\\ &\stackrel{c}{=} \frac{\prod_{i=n-k}^{n}(1-q^i)}{\prod_{i=1}^{k+1}(1-q^i)}\end{aligned}$$

where equality *a* follows from (22), equality *b* follows from the induction hypothesis, and equality *c* follows from Proposition 12.

■

Proposition 12

$$\frac{1}{q^{k+1}} \sum_{j=k+1}^{n} q^j \prod_{i=1}^{k} (1-q^{j-i}) = \frac{1}{1-q^{k+1}} \prod_{i=n-k}^{n} (1-q^i) \qquad \forall n > k.$$

Proof We prove by mathematical induction on n.

(a) *Base Case*: Carefully substituting $n = k+1$ in both the left hand side and the right hand side yields equality.

(b) *Inductive step*:

$$\begin{aligned}
\frac{1}{q^{k+1}} \sum_{j=k+1}^{n+1} q^j \prod_{i=1}^{k} (1-q^{j-i}) &= \frac{1}{q^{k+1}} \left(\sum_{j=k+1}^{n} q^j \prod_{i=1}^{k} (1-q^{j-i}) + q^{n+1} \prod_{i=n-k+1}^{n} (1-q^i) \right) \\
&= \frac{1}{1-q^{k+1}} \prod_{i=n-k}^{n} (1-q^i) + q^{n-k} \prod_{i=n-k+1}^{n} (1-q^i) \\
&= \left(\frac{1-q^{n-k}}{1-q^{k+1}} + q^{n-k} \right) \prod_{i=n-k+1}^{n} (1-q^i) \\
&= \frac{\prod_{i=n-k+1}^{n+1} (1-q^i)}{1-q^{k+1}}
\end{aligned}$$

where in the second equality we used the induction hypothesis.

■

References

W. S. Cleveland. *The Elements of Graphing Data*. Wadsworth Publ. Co., 1985.

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

P. Diaconis. *Group Representations in Probability and Statistics*, volume 11 of *IMS Lecture Notes – Monograph Series*. Institute of Mathematical Statistics, 1988.

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

M. A. Fligner and J. S. Verducci. Distance based ranking models. *Journal of the Royal Statistical Society B*, 43:359–369, 1986.

M. A. Fligner and J. S. Verducci. Multistage ranking models. *Journal of the American Statistical Association*, 83:892–901, 1988.

M. A. Fligner and J. S. Verducci, editors. *Probability Models and Statistical Analyses for Ranking Data*. Springer-Verlag, 1993.

J. Huang, C. Guestrin, and L. Guibas. Efficient inference for distributions on permutations. In *Advances in Neural Information Processing Systems 20*, pages 697–704. 2008.

M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30, 1938.

R. Kondor, A. Howard, and T. Jebara. Multi-object tracking with representations of the symmetric group. In *Artificial Intelligence and Statistics*, 2007.

G. Lebanon and J. Lafferty. Conditional models on the ranking poset. In *Advances in Neural Information Processing Systems, 15*, 2003.

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

J. I. Marden. *Analyzing and Modeling Rank Data*. CRC Press, 1996.

P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: an open architecture for collaborative filtering of netnews. In *Proceedings of the Conference on Computer Supported Cooperative Work*, 1994.

R. P. Stanley. *Enumerative Combinatorics*, volume 1. Cambridge University Press, 2000.

G. L. Thompson. Generalized permutation polytopes and exploratory graphical methods for ranked data. *The Annals of Statistics*, 21(3):1401–1430, 1993.

M. P. Wand and M. C. Jones. *Kernel Smoothing*. Chapman and Hall/CRC, 1995.