Journal of Machine Learning Research 11 (2010) 3481-3518 Submitted 6/09; Revised 8/10; Published 12/10

An Exponential Model for Infinite Rankings

Marina Meilă MMP@STAT.WASHINGTON.EDU
Le Bao LEBAO@STAT.WASHINGTON.EDU
Department of Statistics
University of Washington
Seattle, WA 98195-4322, USA

Editor: David Blei

Abstract

This paper presents a statistical model for expressing preferences through rankings, when the number of alternatives (items to rank) is large. A human ranker will then typically rank only the most preferred items, and may not even examine the whole set of items, or know how many they are. Similarly, a user presented with the ranked output of a search engine, will only consider the highest ranked items. We model such situations by introducing a stagewise ranking model that operates with finite ordered lists called top-t orderings over an infinite space of items. We give algorithms to estimate this model from data, and demonstrate that it has sufficient statistics, being thus an exponential family model with continuous and discrete parameters. We describe its conjugate prior and other statistical properties. Then, we extend the estimation problem to multimodal data by introducing an *Exponential-Blurring-Mean-Shift* nonparametric clustering algorithm. The experiments highlight the properties of our model and demonstrate that infinite models over permutations can be simple, elegant and practical.

Keywords: permutations, partial orderings, Mallows model, distance based ranking model, exponential family, non-parametric clustering, branch-and-bound

1. Introduction

The stagewise ranking model of Fligner and Verducci (1986), also known as *generalized Mallows (GM)*, has been recognized as particularly appropriate for modeling the human process of ranking. This model assigns a permutation π over n items a probability that decays exponentially with its distance to a *central permutation* σ. Here we study this class of models in the limit $n \to \infty$, with the assumption that out of the infinitely many items ordered, one only observes those occupying the first t ranks.

Ordering an infinite number of items is common in retrieval tasks: search engines, programs that match a face, or a fingerprint, or a biological sequence against a database, all output the first t items in a ordering over virtually infinitely many objects. We shall call this output a *top-t* ordering. Unlike machines, people can only reliably rank a small number of items. The GM model has been successfully used to model human ranking decisions. We can view the difference between the standard GM model and the *Infinite GM model* that we introduce here as the difference between an election where each voter returns an ordering of a small number of preselected candidates (nominees) and a “grassroots” election process, where everyone can nominate and order their own favourites from a virtually unlimited population. For instance, the difference between “Order the following issues by how much you care about them” vs. “List in order the issues that you care most about” illustrates the difference between the standard and the Infinite GM models. By these examples, we argue that the Infinite GM corresponds to realistic scenarios. An even more realistic scenario, that we will not tackle for now, is one where a voter (ranker) does not have access to the whole population of items (e.g., a search engine only orders a subset of the web, or a reviewer only evaluates a subset of the submissions to a conference).

After defining the infinite GM model, we show that it has sufficient statistics and give algorithms for estimating its parameters from data in the Maximum Likelihood (ML) framework. To be noted that our model will have an infinite number of parameters, of which only a finite number will be constrained by the data from any finite sample.

Then, we consider the non-parametric clustering of top-t ranking data. Non-parametric clustering is motivated by the fact that in many real applications the number of clusters is not known and outliers are possible. Outliers are known to throw off estimation in an exponential model, unless the tails are very heavy. We introduce an adapted version of the well known Gaussian Blurring Mean-Shift algorithm (Carreira-Perpiñán, 2006) (GBMS) that we call exponential blurring Mean Shift (EBMS).

2. The Infinite Generalized Mallows Model

In this section, we give definitions of key terms used in the article and introduce the *Infinite Generalized Mallows* (IGM) model.

2.1 Permutations, Infinite Permutations and top-t Orderings

A permutation σ is a function from a set of *items* $\{i_1, i_2, \ldots i_n\}$ to the set of *ranks* $1:n$. W.l.o.g. the set of items can be considered to be the set $1:n$. Therefore $\sigma(i)$ denotes the *rank* of item i and $\sigma^{-1}(j)$ denotes the item at rank j in σ.

There are many other ways to represent permutations, of which we will use three, the *ranked list*, the *matrix* and the *inversion table* representation; all three will be defined shortly.

In this paper, we consider permutations over a countable set of items, assumed for convenience to be the the set of positive natural numbers $\mathbb{P} = \{1, 2, \ldots, i \ldots\}$. It is easy to see that the notations $\sigma(i), \sigma^{-1}(j)$ extend immediately to countable items. This will be the case with the other definitions; hence, from now on, we will always consider that the set of items is $\mathbb{P}$.

Any permutation σ can be represented by the (infinite) *ranked list* $(\sigma^{-1}(1)|\sigma^{-1}(2)|\ldots|\sigma^{-1}(j)|\ldots)$. For example, let $\sigma = (2|3|1|5|6|4|\ldots|3n-1|3n|3n-2|\ldots)$. Then $\sigma(1) = 3$ means that item 1 has rank 3 in this permutation; $\sigma(2) = 1$ means that item 2 has rank 1, etc. Conversely, $\sigma^{-1}(1) = 2$ and $\sigma^{-1}(3j) = 3j-2$ mean that the first in the list representation of σ is item 2, and that at rank $3j$ is to be found item $3j-2$, respectively. Often we will call the list representation of a permutation an *ordering*.

A top-t ordering π is the prefix $(\pi^{-1}(1)|\pi^{-1}(2)|\ldots|\pi^{-1}(t))$ of some infinite ordering. For instance, the top-3 ordering of the above σ is $(2|3|1)$.

A top-t ordering can be seen as defining a set consisting of those infinite orderings which start with the prefix π. If we denote by $\mathcal{S}_{\mathbb{P}}$ the set of all permutations over $\mathbb{P}$ and by $\mathcal{S}_{\mathbb{P}-t} = \{\sigma \in \mathcal{S}_{\mathbb{P}} \,|\, \sigma(i) = i, \text{for}\, i = 1:t\}$ the subgroup of all permutations that leave the top-t ranks unchanged, then a top-t ordering π corresponds to a unique element of the right coset $\mathcal{S}_{\mathbb{P}}/\mathcal{S}_{\mathbb{P}-t}$.

We will use Greek letters like π and σ for both full permutations and for top-t orderings to keep the notation light. But we will distinguish almost always between "central permutations" ideal infinite objects denoted by σ, and observed orderings, denoted by π, which by virtue of being observed, are always top-t , that is, truncated. Hence, unless otherwise stated, π will denote a top-t ordering, while σ will denote an infinite permutation.

2.2 The Permutation Matrix Representation and the Inversion Table

Now we introduce the two other ways use to represent permutations and top-t orderings.

For any σ, the *permutation matrix* Σ corresponding to σ has $\Sigma_{ij} = 1$ iff $\sigma(i) = j$ and $\Sigma_{ij} = 0$ otherwise. If σ is an infinite permutation, Σ will be an infinite matrix with exactly one 1 in every row and column. For two permutations σ, σ' over $\mathbb{P}$, the matrix product $\Sigma\Sigma'$ corresponds to the function composition $\sigma' \circ \sigma$.

The matrix Π of a top-t ordering π is a truncation of some infinite permutation matrix Σ. It has t columns, each with a single 1 in $\pi^{-1}(j)$, for $j = 1 : t$.

For example, if $\sigma = (2|3|1|7|4|\ldots)$ and $\pi = (2|3|1)$ is its top-3 ordering, then

$$\Sigma = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & \ldots \\ 1 & 0 & 0 & 0 & 0 & \ldots \\ 0 & 1 & 0 & 0 & 0 & \ldots \\ 0 & 0 & 0 & 0 & 1 & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \end{bmatrix} \quad \text{and} \quad \Pi = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ \ldots & \ldots & \ldots \end{bmatrix}. \qquad (1)$$

For a permutation σ and a top-t ordering π, the matrix $\Sigma^T \Pi$ corresponds to the list of ranks in σ of the items in π. In this context, one can consider σ as a one-to-one relabeling of the set $\mathbb{P}$.

The *inversion table* of a permutation σ, with respect to the identity permutation id is an infinite sequence of non-negative integers $(s_1, s_2, \ldots)$ which are best defined algorithmically and recursively. We consider the ranked list $(1|2|3|\ldots)$. In it, we find $\sigma^{-1}(1)$ the first item of σ, and count how many positions past the head of the list it lies. This is s_1, and it always equals $\sigma^{-1}(1) - 1$. Then we delete the entry $\sigma^{-1}(1)$ from the list, and look up $\sigma^{-1}(2)$; s_2 is the number of positions past the head of this list where we find $\sigma^{-1}(2)$. We then delete $\sigma^{-1}(2)$ as well and proceed to find $\sigma^{-1}(3)$, which will give us s_3, etc. By induction, it follows that an infinite permutation can be represented uniquely by the list $(s_1, s_2, \ldots)$. Hence $s_j \in \{0, 1, 2, \ldots\}$, and, denoting by $1_{[p]}$ the function which is 1 if the predicate p is true and is 0 otherwise, we have

$$s_j(\sigma) \;=\; \sigma^{-1}(j) - 1 - \sum_{j' < j} 1_{[\sigma^{-1}(j') < \sigma^{-1}(j)]}.$$

It is also easy to see that, if π is a top-t ordering, it can be uniquely represented by an inversion table of the form $(s_1, \ldots s_t)$. If π is the top-t ordering of an infinite permutation σ, then the inversion table of π is the t-prefix of the inversion table of σ. This property makes the inversion table particularly convenient for our purposes.

For example, if $\sigma = (2|3|1|7|4|\ldots)$ and $\pi = (2|3|1)$, then $s_1(\sigma) = s_1(\pi) = 1$, $s_2(\sigma) = s_2(\pi) = 1$, $s_3(\sigma) = s_3(\pi) = 0$, $s_4(\sigma) = 3$, $s_5(\sigma) = 0$, etc.

The inversion table has a particularly simple interpretation in the matrix representation of σ or π: s_1 equals the number of zeros preceding 1 in column 1; we delete the row containing this 1, then count the number of zeros in column 2 preceding the 1 to obtain s_2; we delete the row containing this 1, then go to column 3 to count the zeros preceding the 1 in column 3 in order to obtain s_3, and so on. The reader can verify that $s_{1:3}(\pi) = (1, 1, 0)$ from the matrix Π in Equation (1) above.

Another property of the inversion table is that it can be defined with respect to any infinite permutation σ_0, by letting the ordered list corresponding to σ_0 replace the list $(1|2|3|\ldots)$ in the above definition of the inversion table as follows:

$$s_j(\sigma|\sigma_0) \;=\; \sigma_0(\sigma^{-1}(j)) - 1 - \sum_{j'<j} 1_{[\sigma_0(\sigma^{-1}(j'))<\sigma_0(\sigma^{-1}(j))]}. \tag{2}$$

In other words, $1+s_j$ is the rank of $\sigma^{-1}(j)$ in $\sigma_0\left|_{\mathbb{P}\setminus\{\sigma^{-1}(1),\ldots\sigma^{-1}(j-1)\}}\right.$.

In matrix representation, $s_1(\sigma|\sigma_0)$ is the number of 0's preceding the 1 in the first column of $\Sigma_0^T\Sigma$; after we delete the row containing this 1, $s_2(\sigma|\sigma_0)$ is the number of 0's preceding the 1 in the second column, and so on. For a top-t ordering π, this operation is done on the matrix $\Sigma_0^T\Pi$.

For example, for $\pi = (3|2|1)$ and $\sigma_0 = (3|4|2|1|\ldots)$ the matrix representation is

$$\Sigma_0^T\Pi \;=\; \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \cdots & \cdots & \cdots \end{bmatrix},$$

and $s_1(\pi|\sigma_0) = 0, s_2(\pi|\sigma_0) = 1, s_3(\pi|\sigma_0) = 1$.

If σ_0 is given, π is completely determined by the inversion table $s_{1:t}(\pi|\sigma)$. Equation (2) can be interpreted as a recursive algorithm to construct π from σ, which we briefly describe here. We imagine σ to be an ordered list of available items. From it, we choose the first rank in π by skipping the first s_1 ranks in σ and picking $\pi^{-1}(1) = \sigma^{-1}(s_1+1)$. Once $\pi^{-1}(1)$ is picked, this item is deleted from the ordering σ. From this new list of available items, the second rank in π is picked by skipping the first s_2 ranks, and chosing the item in the (s_2+1)-th rank. This item is also deleted, and one proceeds to choose $\pi^{-1}(3), \pi^{-1}(4), \ldots \pi^{-1}(t)$, etc. in a similar manner. This reconstruction algorithm proves that the representation $(s_{1:t})$ uniquely determines π. It is also easy to see that if π is a prefix of σ, that is, if $\pi^{-1}(j) = \sigma^{-1}(j)$ for $j = 1:t$, then $s_1 = s_2 = \ldots = s_t = 0$.

For an example we now show how to reconstruct $\pi = (3|2|1)$ using $\sigma = (3|4|2|1|\ldots)$. Recall that the inversion table of π is given by $s_1(\pi|\sigma) = 0, s_2(\pi|\sigma) = 1, s_3(\pi|\sigma) = 1$.

Stage	π	σ	Comments
Initial	$()$	$(3|4|2|1|\ldots)$	
$j=1$	(3)	$(\cancel{3}|4|2|1|\ldots)$	Skip $s_1 = 0$ ranks from σ, then assign the current item to $\pi^{-1}(1)$
$j=2$	$(3|2)$	$(\cancel{3}|4|\cancel{2}|1|\ldots)$	Skip $s_2 = 1$ ranks from σ, then assign the current item to $\pi^{-1}(2)$
$j=3$	$(3|2|1)$	$(\cancel{3}|4|\cancel{2}|\cancel{1}|\ldots)$	Skip $s_3 = 1$ ranks from σ, then assign the current item to $\pi^{-1}(3)$

The definition of the inversion table s is identical to the first equation of Section 3 from Fligner and Verducci (1986). A reciprocal definition of the inversion table is used by Meilă et al. (2007) and Stanley (1997) and is typically denoted by $(V_1, V_2, \ldots)$. The "V" form of the inversion table is closely related to the inversion table we use here. We discuss this relationship in Section 7.

2.3 Kendall Type Divergences

For finite permutations of n items π and σ,

$$d_K(\pi,\sigma) \;=\; \sum_{j=1}^{n-1} s_j(\pi|\sigma) \;=\; \sum_{j=1}^{n-1} s_j(\sigma|\pi)$$

denotes the *Kendall distance* (Mallows, 1957) (or *inversion distance*) which is a metric. In the above, the index j runs only to $n-1$ because for a finite permutation, $s_n \equiv 0$. The Kendall distance represents the number of adjacent transpositions needed to turn π into σ. An extension of the Kendall distance which has been found very useful for modeling purposes was introduced by Fligner and Verducci (1986). It is

$$d_\theta(\pi,\sigma) \;=\; \sum_{j=1}^{n-1} \theta_j s_j(\pi|\sigma),$$

with $\theta = (\theta_1, \ldots \theta_{n-1})$ a vector of real parameters, typically non-negative. Note that d_θ is in general not symmetric, nor does it always satisfy the triangle inequality.

For the case of countable items, we introduce the divergence

$$d_\theta(\pi,\sigma) \;=\; \sum_{j=1}^{t} \theta_j s_j(\pi|\sigma), \tag{3}$$

where π is a top-t ordering, σ is a permutation in $\mathcal{S}_{\mathbb{P}}$, and $\theta = (\theta_{1:t})$ a vector of *strictly positive* parameters.[1]

When θ_j are all equal $d_\theta(\pi,\sigma)$ is proportional to the Kendall distance between σ and the set of orderings compatible with π and counts the number of inversions needed to make σ compatible with π. In general, this "distance" between a top-t ordering and an infinite ordering is a *set distance*.[2]

2.4 A Probability Model over top-t Rankings of Infinite Permutations

Now we are ready to introduce the *Infinite Generalized Mallows* (IGM) model. We start with the observation that as any top-t ordering can be represented uniquely by a sequence of t natural numbers, defining a distribution over the former is equivalent to defining a distribution over the latter, which is a more intuitive task. In keeping with the GM paradigm of Fligner and Verducci (1986), each s_j is sampled independently from a discrete exponential with parameter $\theta_j > 0$.

$$P(s_j) \;=\; \frac{1}{\psi(\theta_j)} e^{-\theta_j s_j}, \qquad s_j \;=\; 0, 1, 2, \ldots. \tag{4}$$

The normalization constant ψ is

$$\psi(\theta_j) = \sum_{k=0}^{\infty} e^{-\theta_j k} = \frac{1}{1-e^{-\theta_j}}, \tag{5}$$

and the expectation of s_j is $E[s_j|\theta_j] = \frac{e^{-\theta_j}}{1-e^{-\theta_j}} = \frac{1}{e^{\theta_j}-1}$ the well known expectation of the discrete geometric distribution. Now we fix a permutation σ. Since any π is uniquely defined by σ and the inversion table $s_{1:t}(\pi|\sigma)$, Equations (4) and (5) define a distribution over top-t orderings, by $P_{\theta,\sigma}(\pi) = \prod_{j=1}^{t} P(s_j(\pi|\sigma))$. This is equivalent to

$$P_{\theta,\sigma}(\pi) \;=\; e^{-\sum_{j=1}^{t}[\theta_j s_j(\pi|\sigma)+\ln\psi(\theta_j)]}. \tag{6}$$

[1] Definition 3 can be easily extended to a pair $\sigma,\sigma' \in \mathcal{S}_{\mathbb{P}}$, but in this case the divergence will often take infinite values.

[2] A set distance, often called "distance" between two sets, is the minimum distance between elements in different sets, that is, $\delta(A,B) = \min_{x\in A, y\in B} \delta(x,y)$ for a metric or divergence δ. The set distance is not a metric, as it can easily be shown by counterexample.

The above $P_{\theta,\sigma}(\pi)$ has a t-dimensional real parameter θ and an infinite-dimensional discrete parameter σ. The normalization constant $\prod_{j=1}^{t}\psi(\theta_j)$ ensures that

$$\sum_{\pi\in\text{top-}t\text{ orderings of }\mathbb{P}} P_{\theta,\sigma}(\pi) \;=\; 1.$$

In contrast with the finite GM, the parameters θ_j must be strictly positive for the probability distribution to exist. The most probable π for any given t has $s_1(\pi|\sigma) = \ldots = s_t(\pi|\sigma) = 0$. This is the top-$t$ prefix of σ.

The permutation σ is called the *central permutation* of $P_{\theta,\sigma}$. The parameters θ control the spread around the mode σ. Larger θ_j correspond to more concentrated distributions. These facts are direct extensions of statements about the GM model from Fligner and Verducci (1986) and therefore the detailed proofs are omitted.

What is different about the IGM model definition w.r.t its finite counterpart is that the parameter σ is now an infinite sequence instead of a finite one. Another difference is the added condition that $\theta_j > 0$ which ensures that $\psi(\theta_j)$ is finite. This condition is not necessary in the finite case, which leads to the non-identifiability[3] of the parameter σ.

If $\theta_1 = \theta_2 = \ldots = \theta$ the IGM model is called a *single parameter* IGM model. In this case Equation (6) simplifies to

$$P_{\theta,\sigma}(\pi) \;=\; e^{-\theta\sum_{j=1}^{t} s_j(\pi|\sigma) - t\ln\psi(\theta)}.$$

2.5 The IGM Model is a Marginal Distribution

Any top-t ordering π stands for a set of infinite sequences starting with $s_{1:t}(\pi|\sigma)$. Therefore, $P_{\theta,\sigma}(\pi)$ can be viewed as the marginal of $s_{1:t}$ in the infinite product space defined by the distribution

$$P_{\theta,\sigma}(s) \;=\; e^{-\sum_{j=1}^{\infty}[\theta_j s_j + \ln\psi(\theta_j)]} \qquad s\in\mathbb{N}\times\mathbb{N}\times\ldots.$$

Because every infinite sequence s uniquely determines an infinite permutation, the distribution (6) also represents the probability of the π element of the right coset $\mathcal{S}_{\mathbb{P}}/\mathcal{S}_{\mathbb{P}-t}$, that is, the set of infinite permutations that have $(\pi^{-1}(1)|\pi^{-1}(2)|\ldots|\pi^{-1}(t))$ as a prefix. This fact was noted by Fligner and Verducci (1986) in the context of finite number of items. Thus, the IGM model (6) is the infinite counter part of the GM model.

Note also that the expected value of s_j is the mean of the geometric distribution $E[s_j] \;=\; \frac{1}{e^{\theta_j}-1} \;\equiv\; \xi_j$. Thus, the mean value parametrization of the IGM can be easily derived (proof omitted):

$$P_{\xi,\sigma}(\pi) \;=\; \prod_{j=1}^{t} \frac{\xi_j^{s_j(\pi|\sigma)}}{(\xi_j+1)^{s_j(\pi|\sigma)+1}}. \tag{7}$$

It is clear by now that the IGM $P_{\theta,\sigma}$ is an exponential family model over the sample space $(s_1, s_2, \ldots)$ (note that here σ plays no role). It is also evidently an exponential family model in θ over complete or top-t permutations. The next section will demonstrate that, less evidently, the IGM is in fact in the exponential family with respect to the discrete parameter σ as well.

[3] The non-identifiability of the GM model is however not a severe problem for estimation, and can be removed by imposing $\theta_j \geq 0$ (Meilă et al., 2007).

3. Estimating the Model From Data

We are given a set of N top-t orderings $\mathcal{S}_N$. Each $\pi \in \mathcal{S}_N$ can have a different length t_π; all π are sampled independently from a $P_{\theta,\sigma}$ with unknown parameters. We propose to estimate θ, σ from this data in the ML paradigm. We will start by rewriting the log-likelihood of the model, in a way that will uncover a set of sufficient statistics. Then we will show how to estimate the model based on the sufficient statistics.

3.1 Sufficient Statistics

For any square (infinite) matrix $A \in \mathbb{R}^{\mathbb{P}\times\mathbb{P}}$, denote by $L(A) = \sum_{i>j} A_{ij}$ the sum of the elements below the diagonal of A. Let $L_\sigma(A) = L(\Sigma^T A \Sigma)$, and let $\mathbf{1}$ be a vector of all 1's. For any π, let t_π be its length and denote $t_{max} = \max_{\mathcal{S}_N} t_\pi$, $T = \sum_{\mathcal{S}_N} t_\pi$, and by n the number of distinct items observed in the data. As we shall see, t_{max} is the dimension of the concentration parameter θ, n is the order of the estimated central permutation σ, and T counts the total number of items in the data, playing a role akin to the sample size.

Proposition 1 *Let $\{N_j, q_j, Q_j\}_{j\geq 0}$ represent the following statistics: N_j is the number of $\pi \in \mathcal{S}_N$ that have length $t_\pi \geq j$ (in other words, that contain rank j); q_j is the vector $[q_{i,j}]_{i\in\mathbb{P}}$, with $q_{i,j}$ being the number of times i is observed in rank j in the data $\mathcal{S}_N$, $Q_j = [Q_{ii',j}]_{i',i\in\mathbb{P}}$ is a matrix whose element $Q_{ii',j}$ counts how many times $\pi(i) = j$ and $\pi(i') < j$. Then,*

$$\ln P_{\theta,\sigma}(\mathcal{S}_N) \;=\; -\sum_{j\geq 1} [\theta_j L_\sigma(R_j) + N_j \ln\psi(\theta_j)] \quad \text{with}\, R_j \;=\; q_j \mathbf{1}^T - Q_j, \tag{8}$$

To prove this result, we first introduce an alternative expression for the inversion table $s_j(\pi|\sigma)$. Let the data set $\mathcal{S}_N$ consist of a single permutation π and define $q_j(\pi), Q_j(\pi)$ and $R_j(\pi)$ similar to q_j, Q_j, R_j above. Then we have

Proposition 2

$$s_j(\pi|\sigma) \;=\; L_\sigma(q_j(\pi)\mathbf{1}^T - Q_j(\pi)). \tag{9}$$

Proof Let Q_0 be the infinite matrix that has 1 above the main diagonal and 0 elsewhere, $(Q_0)_{ij} = 1$ iff $j > i$ and let $\Pi_{:j}$ denote the j-th column of Π. It is then obvious that $L(A) = \operatorname{trace}(Q_0 A)$ for any A.

By definition, s_j represents the number of 0's preceding 1 in column j, minus all the 1's in the submatrix $(\Sigma^T\Pi)_{1:\sigma(\pi^{-1}(j))-1,1:j-1}$. In other words,

$$\begin{aligned}
s_j(\pi|\sigma) \;&=\; \sum_{l\geq 1} (Q_0\Sigma^T\Pi_{:j})_l(\mathbf{1} - \Sigma^T\Pi_{:1} - \Sigma^T\Pi_{:2} - \ldots \Sigma^T\Pi_{:j-1})_l, \\
&=\; (\mathbf{1} - \sum_{j'<j} \Sigma^T\Pi_{:j'})^T Q_0\Sigma^T\Pi_{:j}, \\
&=\; \mathbf{1}^T Q_0\Sigma^T\Pi_{:j} - \sum_{j'<j} \Pi_{:j'}^T \Sigma Q_0 \Sigma^T \Pi_{:j}, \\
&=\; \operatorname{trace} Q_0\Sigma^T\Pi_{:j}\mathbf{1}^T - \sum_{j'<j} \operatorname{trace} \Pi_{:j}\Pi_{:j'}^T\Sigma Q_0\Sigma^T, \\
&=\; \operatorname{trace} Q_0\Sigma^T[\Pi_{:j}\mathbf{1}^T - \sum_{j'<j} \Pi_{:j}\Pi_{:j'}^T\Sigma], \\
&=\; L(\Sigma^T[\Pi_{:j}\mathbf{1}^T - \sum_{j'<j} \Pi_{:j}\Pi_{:j'}^T]\Sigma) \;=\; L_\sigma(\Pi_{:j}\mathbf{1}^T - \sum_{j'<j} \Pi_{:j}\Pi_{:j'}^T).
\end{aligned}$$

We use the fact that multiplying left by Q_0 counts the zeros preceding 1 in a column in the first equality, $\operatorname{trace} AB = \operatorname{trace} BA$ in the fourth and fifth equations, and the identity $\mathbf{1}^T\Sigma = \mathbf{1}^T$ in the last equation. We now note that $\Pi_{:j} = q_j(\pi)$ and $\sum_{j'<j}\Pi_{:j}\Pi_{:j'}^T = Q_j(\pi)$ and the result follows. □

Proof of Proposition 1. The log-likelihood of $\mathcal{S}_N$ is given by

$$\begin{aligned}\ln P_{\theta,\sigma}(\mathcal{S}_N) &= -\sum_{\pi\in\mathcal{S}_N}\left[\sum_{j=1}^{t}\pi\theta_j s_j(\pi|\sigma) + \ln\psi(\theta_j)\right],\\ &= -\sum_{j\geq 1}\left[\theta_j\sum_{\pi\in\mathcal{S}_N}s_j(\pi|\sigma) + N_j\ln\psi(\theta_j)\right].\end{aligned}$$

Because L_σ is a linear operator, the sum over $\pi\in\mathcal{S}_N$ equals

$$L_\sigma[(\sum_{\pi\in\mathcal{S}_N}q_j(\pi))\mathbf{1}^T - \sum_{\pi\in\mathcal{S}_N}Q_j(\pi))].$$

It is easy to verify now that the first sum represents q_j and the second one represents Q_j.
□

The sufficient statistics for the single parameter IGM model are described by the following corollary.

Corollary 3 *Denote*

$$q = \sum_j q_j, \quad Q = \sum_j Q_j, \quad R \ = \ q\mathbf{1}^T - Q. \tag{10}$$

If $\theta_1 = \theta_2 = \ldots = \theta$ *then the log-likelihood of the data* $\mathcal{S}_N$ *can be written as*

$$\ln P_{\theta,\sigma}(\mathcal{S}_N) \ = \ -\theta L_\sigma(R) - T\ln\psi(\theta). \tag{11}$$

Note that q_i, $Q_{ii'}$ represent respectively the number of times item i is observed in the data and the number of times item i' precedes i in the data.

Proposition 1 and Corollary 3 show that the infinite model $P_{\theta,\sigma}$ has *sufficient statistics*. The result is obtained without any assumptions on the lengths of the observed permutations. The data $\pi\in\mathcal{S}_N$ can have different lengths t_π, t_π may be unbounded, and may even be infinite.

As the parameters θ, σ of the model are infinite, the sufficient statistics R_j (or R) are infinite matrices. However, for any practically observed data set, t_{max} will be finite and the total number of items observed will be finite. Thus, N_j, R_j will be 0 for any $j > t_{max}$ and $R = \sum_{j\in\mathbb{P}} R_j$ will have non-zero entries only for the items observed in some $\pi\in\mathcal{S}_N$. Moreover, with a suitable relabeling of the observed items, one can reduce R_j to a matrix of dimension n, the number of distinct observed items. The rest of the rows and columns of R_j will be 0 and can be discarded. So in what follows we will assume that R_j and the other sufficient statistics have dimension n.

3.2 ML Estimation: The Case of a Single θ

We now go on to estimate θ and σ starting with the case of equal θ_j, that is, $\theta_1 = \theta_2 = \ldots = \theta$. In this case, Equation (11) shows that the estimation of θ and σ decouple. For any fixed σ, Equation (11) attains its maximum over θ at

$$\theta = \ln(1 + T/L_\sigma(R)). \tag{12}$$

In contrast to the above explicit formula, for the finite GM, the likelihood has no analytic solution for θ (Fligner and Verducci, 1986). The estimated value of θ increases when $L_\sigma(R)$ decreases. This has an intuitive interpretation. The lower triangle of $\Sigma^T R\Sigma$ counts the "out of order" events w.r.t. the chosen model σ. Thus, $L_\sigma(R)$ can be seen as a residual, which is normalized by the "sample size" T. Equation (12) can be re-written as $L_\sigma(R)/T = 1/(e^\theta - 1) \equiv \xi$. In other words, the mean value parameter ξ equals the average residual. This recovers a well-known fact about exponential family models. In particular, if the residual is small, that is $L_\sigma(R)$ has low counts, we conclude that the distribution is concentrated, hence has a high θ.

Estimating σ^{ML} amounts to minimizing $L_\sigma(R)$ w.r.t σ, independently of the value of θ. The optimal σ according to Corollary 3 is the permutation that minimizes the lower triangular part of $\Sigma^T R\Sigma$. To find it we exploit an idea first introduced in Meilă et al. (2007). This idea is to search for $\sigma = (i_1|i_2|i_3|\ldots)$ in a stepwise fashion, starting from the top item i_1 and continuing down.

Assuming for a moment that $\sigma = (i_1|i_2|i_3|\ldots)$ is known, the cost to be minimized $L_\sigma(R)$ can be decomposed columnwise as

$$L_\sigma(R) \;=\; \sum_{l\neq i_1} R_{li_1} + \sum_{l\neq i_1,i_2} R_{li_2} + \sum_{l\neq i_1,i_2,i_3} R_{li_3} + \ldots,$$

where the number of non-trivial terms is one less than the dimension of R. It is on this decomposition that the search algorithm is based. Reading the above algorithmically, we can compute $L_\sigma(R)$, for any given σ by the following steps

1. zero out the diagonal of R
2. sum over column i_1 of the resulting matrix
3. remove row and column i_1
4. repeat recursively from step 2 for $i_2, i_3, \ldots$.

If now σ is not known, the above steps 2, 3 become the components of a search algorithm, which works by trying every i_1 in turn, saving the partial sums, then continuing down for a promising i_1 value to try all i_2's that could follow it, etc. This type of search is represented by a *search tree*, whose nodes are candidate prefixes for σ.

The search tree has $n!$ nodes, one for each possible ordering of the observed items. Finding the lowest cost path through the tree is equivalent to minimizing $L_\sigma(R)$. *Branch-and-bound (BB)* (Pearl, 1984) algorithms are methods to explore the tree nodes in a way that guarantees that the optimum is found, even though the algorithm may not visit all the nodes in the tree. The number of nodes explored in the search for σ^{ML} depends on the individual sufficient statistics matrix R. It was shown by Meilă et al. (2007) that in the worst case, the number of nodes searched can be a significant fraction of $n!$ and as such intractable for all but small n. However, if the data distribution is concentrated around a mode, then the search becomes tractable. The more concentrated the data, the more effective the search.

We call the BB algorithm for estimating σ the SIGMA*, by analogy with the name A^* under which such algorithms are sometimes known. The algorithm is outlined in Figure 1. In this figure, A is an *admissible heuristic*; Pearl (1984) explains their role. By default, one can use $A \equiv 0$. A higher bound than 0 will accelerate the search; some of the admissible heuristics of Mandhani and Meilă (2009) can be used for this purpose.

In addition to this slow but exact algorithm, various heuristic search techniques can be used to explore the search tree of the problem. Two of them which showed good performance for the standard GM model and which transfer immediately to the infinite model are the greedy search (GREEDYR) and the the SORTR heuristic of Fligner and Verducci (1988), both described in Figure 2. The SORTR computes the costs of the first step of BB, then outputs the permutation σ that sorts these costs in increasing order. The algorithm as proposed by Fligner and Verducci (1988) also performs limited search around this σ. For simplicity, this was not included in the pseudocode, but can be implemented easily.

The GREEDYR replaces BB with greedy search on the same sufficient statistics matrix. This algorithm was used by Cohen et al. (1999), where a factor of 2 approximation bound w.r.t. the cost was also shown to hold.

A third heuristic is related to the special case $t = 1$, when each π contains only 1 element. This is the situation of, for example, a search engine returning just the best match to a query. For $t = 1$, as Q is 0, the optimal ordering is the one minimizing $L_\sigma(q\mathbf{1}^T) = [0\ 1\ \dots\ n-1]\Sigma^T q$. This is obviously the ordering that sorts the items in descending order of their frequency q.

In conclusion, to estimate the parameters from data in a single parameter case, one first computes the sufficient statistics, then a prefix of σ is estimated by exact or heuristic methods, and finally, with the obtained ordering of the observed items, one can compute the estimate of θ.

3.3 ML Estimation: The Case of General θ

Maximizing the likelihood of the data $\mathcal{S}_N$ is equivalent, by Proposition 1, with minimizing

$$J(\theta,\sigma) \quad = \quad \sum_j [\theta_j L_\sigma(R_j) + N_j \ln \psi(\theta_j)] \;=\; L_\sigma(\underbrace{\sum_j \theta_j R_j}_{R_\theta}) + \text{function of } \theta. \tag{13}$$

This estimation equation does not decouple w.r.t θ and σ. Minimization is however possible, due to the following two observations. First, for any fixed set of θ_j values, minimization w.r.t σ is possible by the algorithms described in the previous section. Second, for fixed σ, the optimal θ_j parameters can be found analytically by

$$\theta_j = \ln(1 + N_j / L_\sigma(R_j)). \tag{14}$$

The two observations immediately suggest an alternating minimization approach to obtaining θ^{ML}, σ^{ML}. The algorithm is given in Figure 3. For the optimization w.r.t σ exact minimization can be replaced with any algorithm that decreases the r.h.s of (13). As both steps increase the likelihood, the algorithm will stop in a finite number of steps at a local optimum.[4]

3.4 Identifiability and Consistency Results

One remarkable property of the IGM, which is easily noted by examining the likelihood in (8) or (11), is that the data will only constrain a finite number of parameters of the model. The log-likelihood (8) depends only on the parameters $\theta_{1:t_{max}}$. Maximizing likelihood will determine $\theta_{1:t_{max}}$ leaving the other θ_j parameters undetermined.

[4] The reader may have noted that $P_{\theta,\sigma}$ is an exponential family model. For exponential family models with *continuous* parameters over a convex set, the likelihood is log-concave and an iteration like the one presented here would end at the global optimum. For our model, however, the parameter σ is discrete; moreover, the set of σ's forms the vertices of a convex polytope. One can show theoretically and practically that optimizing $L_\sigma(R)$ can have multiple optima and hence one cannot expect to always find a global maximum for the likelihood. However, we suspect that under the conditions that make optimization tractable, that is, a concentrated data distribution, existence of a global optimum may be proved.

```
Algorithm SIGMA*
Input matrix $R \in \mathbb{R}^{n\times n}$ of sufficient statistics
    Initialize
        $S = \{\bar{\sigma}_{\emptyset}\}$, $\bar{\sigma}_{\emptyset}$ =the empty ordering, $j = 0$, $C(\bar{\sigma}_{\emptyset}) = B(\bar{\sigma}_{\emptyset}) = 0$
    Do
    remove $\bar{\sigma} \in \underset{\sigma \in S}{\operatorname{argmin}} B(\bar{\sigma})$ from $S$
    if length($\bar{\sigma}$) $= n$ (Return)
        Output $\bar{\sigma}$, $B(\bar{\sigma}) = C(\bar{\sigma})$ and Stop.
    else (Expand $\bar{\sigma}$)
        for $i_{j+1} \in \{1:n\} \setminus \bar{\sigma}$
          1. create node $\bar{\sigma}' = (i_1|\ldots,|i_j|i_{j+1})$
          2. $v_{j+1}(\bar{\sigma}') \ = \ \sum_{l\in\{1:n\}\setminus\bar{\sigma}'} R_{l i_{j+1}}$
          3. calculate $C(\bar{\sigma}') = C(\bar{\sigma}) + v_{j+1}$, calculate $A(\bar{\sigma}')$
          4. $B(\bar{\sigma}') = C(\bar{\sigma}') + A(\bar{\sigma}')$
          5. store node $(\bar{\sigma}', j+1, C(\bar{\sigma}'), B(\bar{\sigma}'))$ in $S$
```

Figure 1: Algorithm SIGMA* outline. S is the set of nodes to be expanded; $\bar{\sigma} = (i_1|\ldots,|i_j)$ designates a top-j ordering, that is, a node in the tree at level j. The cost of the path $\bar{\sigma}$ is given by $C(\bar{\sigma}) = \sum_{j'=1}^{j} \sum_{l \notin \{i_{1:j'}\}} R_{l i_l}$, and $A(\bar{\sigma})$ is a lower bound on the cost to go from $\bar{\sigma}$, possibly 0. The total estimated cost of node $\bar{\sigma}$ is $B(\bar{\sigma}) = C(\bar{\sigma}) + A(\bar{\sigma})$, which is used to predict which is the most promising path through the tree. In an implementation, node $\bar{\sigma}$ stores: $\bar{\sigma} = (i_1|\ldots,|i_j)$, $j = |\bar{\sigma}|, C(\bar{\sigma}), B(\bar{\sigma})$.

Let n be the number of distinct items observed in the data. From σ, we can estimate at most its restriction to the items observed, that is, the restriction of σ to the set $\bigcup_{\pi\in\mathcal{S}_N}\{\pi^{-1}(1), \pi^{-1}(2), \ldots, \pi^{-1}(t_\pi)\}$. The next proposition shows that the ML estimate will always be a permutation which puts the observed items before any unobserved items.

Proposition 4 *Let $\mathcal{S}_N$ be a sample of top-t orderings, and let σ be a permutation over $\mathbb{P}$ that ranks at least one unobserved item i_0 before an item observed in $\mathcal{S}_N$. Then there exists another permutation $\tilde{\sigma}$ which ranks all observed items before any unobserved items, so that for any parameter vector $(\theta_{1:t})$, $P_{\theta,\sigma}(\mathcal{S}_N) < P_{\theta,\tilde{\sigma}}(\mathcal{S}_N)$.*

Proof For an item i_0 not observed in the data, $q_{i_0,j}, Q_{i_0 i,j}$ and $R_{i_0 i,j}$ are 0, for any $j = 1:t$ and any observed item i. Hence row i_0 in any R_j is zero. Also note that if we switch among each other items that were not observed, there is no effect in the likelihood. Hence, w.l.o.g we assume that i_0 has rank j_0 in σ and is followed by an item i which is observed.

```
Algorithm SORTR

Input matrix $R \in \mathbb{R}^{n\times n}$ of sufficient statistics with 0 diagonal

  1. compute the column sums of $R$, $r_l = \sum_k R_{kl}$, $l = 1:n$

  2. sort $r_l$, $l = 1:n$ in increasing order

Output $\sigma$ the sorting permutation
```

```
Algorithm GREEDYR

Input matrix $R \in \mathbb{R}^{n\times n}$ of sufficient statistics with 0 diagonal

  1. set $V = 1:n$ the set of unused items

  2. Repeat for $j = 1:n-1$

       (a) compute $r_l = \sum_{k\in V} R_{kl}$, $l \in V$ the column sums of a submatrix of $R$
       (b) let $l^* = \text{argmin}_{l\in V} r_l$
       (c) set $\sigma^{-1}(j) = l^*$, $V \leftarrow V \setminus \{l^*\}$

  3. set $\sigma^{-1}(n)$ to the last remaining item in $V$

Output $\sigma$
```

Figure 2: Heuristic algorithms to estimate a central permutation: SORTR and GREEDYR. The elements R_{ii} are never part of any s_j, hence to simplify the code we assume they are set to 0.

```
Algorithm ESTIMATESIGMATHETA

Input Sufficient statistics $R_j, N_j$, $j = 1:t_{max}$
    Initial parameter values $\theta_{1:t_{max}} > 0$

    Iterate until convergence:

       1. Calculate $R_\theta = \sum_j \theta_j R_j$
       2. Find the ordering $\sigma = \text{argmin}_\sigma L_\sigma(R_\theta)$ (exactly by SIGMA* or by heuristics)
       3. Estimate $\theta_j = \ln(1 + N_j / L_\sigma(R_j))$

Output $\sigma$, $\theta_{1:t_{max}}$
```

Figure 3: Algorithm ESTIMATESIGMATHETA.

The idea of the proof is to show that if we switch items i_0 and i the lower triangle of any R_j will not increase, and for at least one R_j it will strictly decrease.

For this, we examine row i of some R_j. We have $Q_{ii_0,j} = 0$ and $Q_{ii_0,j} = q_{i,j}$. Since i is observed then for at least one j we have $R_{ii_0,j} > 0$. Denote by σ' the permutation which is equal to σ except for switching i and i_0. The effect of switching i and i_0 on R_j is to switch elements $R_{ii_0,j}$ and $R_{i_0i,j}$. Since the latter is always 0 and the former is greater or equal to 0, it follows that $L_{\sigma'}(R_j) \leq L_\sigma(R_j)$ for any j, and that the inequality is strict for at least one j. By examining the likelihood expression in Equation (8), we can see that for any positive parameters $\theta_{1:j}$ we have $\ln P_{\theta,\sigma}(\mathcal{S}_N) < \ln P_{\theta,\sigma'}(\mathcal{S}_N)$.

By successive switches like the one described here, we can move all observed items before the unobserved items in a finite number of steps. Let the resulting permutation be $\tilde{\sigma}$. In this process, the likelihood will be strictly increasing at each step, therefore the likelihood of $\tilde{\sigma}$ will be higher than that of σ. □.

In other words σ^{ML} is a permutation of the observed items, followed by the unobserved items in any order. Hence the ordering of the unobserved items is completely non-identifiable (naturally so). But not even the restriction of σ to the observed items is always completely determined. This can be seen by the following example. Assume the data consists of the the two top-t orderings $(a|b|c)$, $(a|b|d)$. Then $(a|b|c|d)$ and $(a|b|d|c)$ are both ML estimates for σ; hence, it would be more accurate to say that the ML estimate of σ is the *partial ordering* $(a|b|\{c,d\})$. The reason σ^{ML} is not unique over a,b,c,d in this example is that the data has no information about the relative ranking c,d, neither directly by observing c,d together in the same π, nor indirectly, via a third item. This situation is likely to occur for the rarely observed items, situated near the ends of the observed π's. Thus this kind of inderterminacy will affect predominantly the last ranks of σ^{ML}. Another kind of indeterminacy can occur when the data is ambiguous w.r.t the ranking of two items c,d, that is, when $R_{cd} = R_{dc} > 0$. This situation can occur at any rank, and will occur more often for values of the θ_j parameters near 0. However, observing more data mitigates this problem. Also, since more observations typically increase the counts more for the first items in σ, this type of indeterminacy is also more likely to occur for the later ranks of σ^{ML}.

Thus, in general, there is a finite set of permutations of the observed items which have equal likelihood. We expect that these permutations will agree more on their first ranks and less in the last ranks. The exact ML estimation algorithm SIGMA* described here will return one of these permutations.

We now discuss the convergence of the parameter estimates to their true values. The IGM model has t real parameters θ_j, $j = 1:t$ and a discrete and infinite dimensional parameter, the central permutation σ. We give partial results on the consistency of the ML estimators, under the assumption that the true model is an IGM model and that t the length of the observed permutations is fixed or bounded.

Before we present the results, we need to make some changes in notation. In this section we will denote by $\hat{q}$, $\hat{R}$, etc the statistics obtained from a sample, normalized by the sample size N. We use q, R, etc for the asymptotic, population based expectation of a statistic under the true model $P_{\mathrm{id},\theta}$ (single parameter or multiple parameters as will be the case). For instance $\hat{q}_{i,j} = \sum_{\pi\in\mathcal{S}_N} q_{i,j}(\pi)/N$ represents the frequency of i appearing in position j in the sample, while $q_{i,j}$ is the probability of this event under $P_{\mathrm{id},\theta}$. For simplicity of notation, the dependence of N is omitted.

We will show that under weak conditions, the statistics of the type q, Q, R converge to their expectations, which in turn will entail convergence of the estimates based on them. The proofs are in Appendix A.1.

Proposition 5 *Let σ be any infinite permutation. If the true model is $P_{\mathrm{id},\theta}$ (multiple parameters) and t is fixed, then* $\lim_{N\to\infty} L_\sigma(\hat{R}_j) = L_\sigma(R_j)$ *for any $j = 1:t$.*

Since we can assume w.l.o.g. that the central permutation of the true model is the identity permutation, this proposition implies that for any IGM, with single or multiple parameters, and for any σ, the statistics $L_\sigma(\hat{R}_j)$ are consistent when t is constant.

Proposition 6 *If the true model is $P_{\mathrm{id},\theta}$ (multiple or single parameter) and t is fixed, then for any infinite permutation σ, denote by $\hat{\theta}_j(\sigma)$ (or $\hat{\theta}(\sigma)$) the ML estimate of θ_j (or θ) given that the estimate of the central permutation is σ. Then for $j = 1:t$*

$$\begin{aligned}\lim_{N\to\infty} \hat{\theta}_j(\sigma) &= \theta_j(\sigma),\\ \lim_{N\to\infty} \hat{\theta}(\sigma) &= \theta(\sigma),\end{aligned}$$

where the limits should be taken in the sense of convergence in probability.

Proposition 7 *Assume that the true model is $P_{\mathrm{id},\theta}$, t is fixed, and $\theta_j \geq \theta_{j+1}$ for $j = 1:t-1$. Let $\sigma \neq \mathrm{id}$ be an infinite permutation. Then, $P[L_{\mathrm{id}}(\hat{R}_j) < L_\sigma(\hat{R}_j)] \to 1$. Consequently, $P[L_{\mathrm{id}}(\hat{R}) < L_\sigma(\hat{R})] \to 1$.*

The consequences of these results are as follows. Assume the true model has a single parameter, and we are estimating a single parameter IGM model. Then, the likelihood of an infinite permutation σ is given by $\hat{R} = \sum_{j=1}^{t} \hat{R}_j$. By Proposition 7, for any σ other than the true one, the likelihood will be lower than the likelihood of the true permutation, except in a vanishingly small set of cases. This result is weaker than ideal, since ideally we would like to prove that the likelihood of the true σ is higher than that of all other permutations simultaneously. We intend to pursue this topic, but to leave the derivation of stronger and results for a further publication.

Proposition 6 shows that, if the correct σ is known, then the θ_j parameters, or alternatively the single θ parameter, are consistent.

In the multiple parameter IGM case, for any fixed θ, the likelihood of σ is given by $\hat{R} = \sum_{j=1}^{t} \theta_j \hat{R}_j$. Hence, by Proposition 7, in this case too, for any given σ different from the true central permutation, the likelihood of σ will be lower than the likelihood of the true model permutation.

The reader will note that these results can be easily extended to the case of bounded t.

4. Non-parametric Clustering

The above estimation algorithms can be thought of as finding a *consensus ordering* for the observed data. When the data have multiple modes, the natural extension to optimizing consensus is clustering, that is, finding the groups of the population that exhibit consensus.

Having defined a distance and a method for estimating ML parameters gives one access to a large number of the existing clustering paradigms originally defined for Euclidean data. For instance, the extensions of the K-means and EM algorithms to infinite orderings is immediate, and so are extensions to other distance-based clustering methods. Here we will present only one clustering method, a the *Exponential Blurring Mean-Shift (EBMS)*, but which illustrates well the issues of clustering in the space of top-t orderings. The EBMS is a nonparametric clustering method.

```
Algorithm EBMS

Input Top-$t$ orderings $\mathcal{S}_N = \{\pi_i\}_{i=1:N}$, with length $t_i$; optionally, a scale parameter $\theta$

    1. For $\pi_i \in \mathcal{S}_N$ compute $q_i, Q_i, R_i$ the sufficient statistics of a single data point.

    2. Reduce data set by counting only the distinct permutations to obtain reduced $\tilde{\mathcal{S}}_N$ and
       counts $N_i \geq 1$ for each ordering $\pi_i \in \tilde{\mathcal{S}}_N$.

    3. For $\pi_i, \pi_j \in \tilde{\mathcal{S}}_N$ calculate Kendall distance $d_{ij} = d_K(\pi_i, \pi_j)$.

    4. (Optional, if $\theta$ not given in input) Set $\theta$ by solving the equation

       $E_\theta[d(\tilde{\mathcal{S}}_N)] = \frac{t_\pi e^{-\theta}}{1-e^{-\theta}} - \sum_{j=1}^{t_\pi} \frac{je^{-\theta}}{1-e^{-j\theta}}$

       where we set $E_\theta[d(\tilde{\mathcal{S}}_N)]$ to be the average of pairwise distances in step 3.

    5. For $\pi_i \in \tilde{\mathcal{S}}_N$        (Compute weights and shift)

        (a) For $\pi_j \in \tilde{\mathcal{S}}_N$: set $\alpha_{ij} = \frac{\exp(-\theta d_{ij})}{\sum_{j'=1}^{n} \exp(-\theta d_{ij'})}$
        (b) Calculate $\overline{R_i} = \sum_{\pi_j \in \tilde{\mathcal{S}}_N} N_j \alpha_{ij} R_j$
        (c) Estimate $\sigma_i$ the "central" permutation that optimizes $\overline{R_i}$
            (exactly or by heuristics)
        (d) Set $\pi_i \leftarrow \sigma_i(1:t_\pi)$
    6. Go to step 2, until no $\pi_i$ changes.
  Output $\tilde{\mathcal{S}}_N$
```

Figure 4: The EBMS algorithm.

Nonparametric clustering is motivated by the fact that in many real applications the number of clusters is unknown and outliers exist. We consider an adapted version of the well known blurring mean-shift algorithm for ranked data (Fukunaga and Hostetler, 1975; Cheng, 1995; Carreira-Perpiñán, 2006). We choose the exponential kernel with bandwidth $\frac{1}{\theta} > 0$: $K_\theta(\pi, \sigma) = \frac{e^{-\theta d(\pi,\sigma)}}{\psi(\theta)}$. Under the Kendall distance $d_K(\pi, \sigma)$ the kernel has the same form as the one parameter Mallows' model. The kernel estimator of π_i is given by

$$\hat{r}(\pi_i) = \sum_{j=1}^{n} \frac{K_\theta(\pi_i, \pi_j)}{\sum_{k=1}^{n} K_\theta(\pi_i, \pi_k)} \pi_j = \sum_{j=1}^{n} \frac{e^{-\theta d_K(\pi_i, \pi_j)}}{\sum_{k=1}^{n} e^{-\theta d_K(\pi_i, \pi_k)}} \pi_j,$$

which does not depend on the normalizing constant $\psi(\theta)$ in Mallows' model.

The EBMS algorithm is summarized in Figure 4. It shift the "points" (i.e., top-t orderings) to new locations obtained by a locally weighted combination of all the data. Thus, every π is "attracted" towards its closest neighbors; as the shifting is iterated the data collapse into one or more clusters. The algorithm has a *scale parameter* θ. The scale influences the size of the local neighborhood of a top-t ordering, and thereby controls the granularity of the final clustering: for small θ values (large neighborhoods), points will coalesce more and few large clusters will form; for large θ's the orderings will cluster into small clusters and singletons. In the EBMS algorithm, we estimate the scale parameter θ at each iteration by solving the equation in step (d).

Practical experience shows that blurring mean-shift merges the points into compact clusters in a few iterations and then these clusters do not change but simply approach each other until they eventually merge into a single point (Carreira-Perpiñán, 2006). Therefore, to obtain a meaningful clustering, a proper stopping criterion should be proposed in advance. For ranked data, this is not the case: since at each iteration step we round the local estimator into the nearest permutation, the algorithm will stop in a finite number of steps, when no ordering moves from its current position. Moreover, because ranked data is in a discrete set, we can also perform an accelerating process. As soon as two or more orderings become identical, we replace this *cluster* with a single ordering with a weight proportional to the cluster's number of members. The total number of iterations remains the same as for the original exponential blurring mean-shift but each iteration uses a data set with fewer elements and is thus faster.

In the algorithm one evaluates distances between top-t orderings. There are several ways in which to turn $d_K(\pi,\sigma)$ into a $d(\pi_1,\pi_2)$, where both terms are top-t orderings, containing different sets of items. Critchlow (1985) studied them, and here we adopt for $d(\pi_1,\pi_2)$ what is called the set distance, that is, the distance between the sets of infinite orderings compatible with π_1 respectively π_2.

We chose this formulation rather than others because this distance equals 0 when $\pi_1=\pi_2$, and this is good, one could even argue necessary, for clustering. In addition, it can be calculated by a relatively simple formula, inspired by Critchlow (1985). Let

A	the set intersection of orderings π_1,π_2	t_1,t_2	lengths of π_1,π_2
B	$\pi_1\setminus A$, the items in π_1 not in π_2	$n_{A,B,C}$	the number of items in A,B,C
C	$\pi_2\setminus A$, the items in π_2 not in π_1	k_j	the index in π_1 of the j-th item not in A
		l_j	the index in π_2 of the j-th item not in A

Then

$$d(\pi_1,\pi_2) \;=\; d_K((\pi_1)_{|A},(\pi_2)_{|A})+n_Bn_C+n_Bt_1-\sum_{j=1}^{n_B}k_j-\frac{n_B(n_B-1)}{2}+n_Ct_2-\sum_{j=1}^{n_C}l_j-\frac{n_C(n_C-1)}{2}. \tag{15}$$

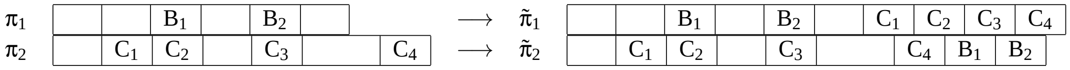


Figure 5: An example of obtaining two partial orderings $\tilde\pi_1,\tilde\pi_2$ compatible respectively with π_1,π_2 that achieve the set distance. Empty spaces represent the common items, while B and C symbols mark the items in B, respectively C. The distance $d(\pi_1,\pi_2)$ is the Kendall distance between $\tilde\pi_1$ and $\tilde\pi_2$.

The intuitive interpretation of this distance is given in Figure 5. We extend π_1 and π_2 to two longer orderings $\tilde{\pi}_1, \tilde{\pi}_2$, so that: (i) $\tilde{\pi}_1$, $\tilde{\pi}_2$ have identical sets of items, (ii) $\tilde{\pi}_1$ is the closest ordering to π_2 which is compatible with π_1, and (iii) reciprocally, $\tilde{\pi}_2$ is the closest ordering to π_1 which is compatible with π_2. We obtain $\tilde{\pi}_1$ by taking all items in π_2 but not in π_1 and appending them at the end of π_1 while preserving their relative order. A similar operation gives us $\tilde{\pi}_2$. Then, $d(\pi_1,\pi_2) = d_K(\tilde{\pi}_1,\tilde{\pi}_2)$, and Equation (15) expresses this value.

5. The Conjugate Prior

The existence of sufficient statistics implies the existence of a conjugate prior (DeGroot, 1975) for the parameters of model (6). Here we introduce the general form of this prior and show that computing with the conjugate prior (or posterior), is significantly harder than computing with the likelihood (6).

We shall assume for simplicity that all top-t rankings have the same t. Consequently, our parameter space consists of the real positive vector $\theta_{1:t}$ and the discrete infinite parameter Σ.

We define the prior parameters as a set of “fictitious sufficient statistics”, by analogy with the sufficient statistics for model (6). For this we first make a few straightforward observations about the sufficient statistics q_j, Q_j, $j = 1:t$ as follows:

$$\begin{aligned}
Q_1 &\equiv 0, \\
Q_{ii',j} &\geq 0 \quad \text{for all } i, i', j, \\
\sum_i q_{i,j} &= N_j \quad \text{for all } j, \\
Q_j \mathbf{1} &= (j-1) q_j \quad \text{for all } j > 1.
\end{aligned}$$

Therefore

$$R_j \;=\; q_j \mathbf{1}^T - Q_j \;=\; \begin{cases} Q_j \left(\frac{1}{j-1} \mathbf{1}\mathbf{1}^T - I\right) & \text{for } j > 1 \\ q_1 \mathbf{1}^T & \text{for } j = 1 \end{cases} .$$

Now we let ν denote the *prior strength*, representing the equivalent sample size, and λ_1, Λ_j, $j = 2:t$ be the prior parameters corresponding to the sufficient statistics q_1, $Q_{2:t}$, normalized as follows.

Proposition 8 *Let $\nu > 0$, λ_1 be a vector and Λ_j, $j = 2:t$ denote a set of possibly infinite matrices satisfying*

$$\begin{aligned}
\lambda_1 &\geq 0, \\
\Lambda_{ii',j} &\geq 0 \quad \text{for all } i, i', j, \\
\Lambda_j \mathbf{1} &= (j-1)\lambda_j \quad \text{for all } j > 1 \quad \text{(by definition)}, \\
\mathbf{1}^T \lambda_j &= 1 \quad \text{for all } j.
\end{aligned}$$

Denote $\mathbf{\Lambda} = \{\nu, \lambda_1, \Lambda_{2:t}\}$ and

$$R_j^0 \;=\; \begin{cases} \Lambda_j \left(\frac{1}{j-1} \mathbf{1}\mathbf{1}^T - I\right) & \text{for } j > 1 \\ \lambda_1 \mathbf{1}^T & \text{for } j = 1 \end{cases} .$$

Define the distribution

$$P_{\mathbf{\Lambda}}(\sigma,\theta) \;\propto\; e^{-\nu \sum_{j=1}^t [\theta_j L(\Sigma^T R_j^0 \Sigma) + \ln \psi(\theta_j)]}, \tag{16}$$

which is a conjugate prior for the model $P_{\theta,\sigma}(\pi)$ defined in (6).

Proof Given observed permutations $\pi_{1:N}$ with sufficient statistics R_j, $j = 1:t$, the posterior distribution of (σ,θ) is updated by

$$\begin{aligned} P(\theta,\sigma\,|\,\mathbf{\Lambda},\pi_{1:N}) &\propto e^{-\sum_{j=1}^{t}[(\nu L_\sigma(R_j^0)+L_\sigma(R_j))\theta_j+(N+\nu)\ln\psi(\theta_j)]}, \\ &= e^{-(N+\nu)\sum_{j=1}^{t}[\theta_j L_\sigma\left(\frac{\nu R_j^0+R_j}{N+\nu}\right)+\ln\psi(\theta_j)]}. \end{aligned}$$

If the hyperparameters $\nu,\lambda_1,\Lambda_{2:t}$ satisfy the conditions of the proposition, then the new hyperparameters $\mathbf{\Lambda}' = \{\nu+N,\,(\nu\lambda_1+q_1)/(\nu+N),\,(\nu\Lambda_j+Q_j)/(\nu+N),\; j=2:t\}$ satisfy the same conditions. □.

The conjugate prior is defined in (16) only up to a normalization constant.[5] As it will be shown below, this normalization constant is not always computable in closed form. Another aspect of conjugacy is that one prefers the conjugate hyperparameters to represent expectations of the sufficient statistics under some $P_{\theta,\sigma}$. The conditions in Proposition 8 are necessary, but not sufficient to ensure this fact.

To simplify the notations, we write

$$S_j^* \;=\; L_\sigma(\nu R_j^0 + R_j). \tag{17}$$

This notation reflects the fact that S_j^* is the counterpart in the posterior of the s_j in the distribution $P_{\theta,\sigma}(\pi)$. If $N=0$, then $S_j^* = \nu L_\sigma(R_j^0)$. The value of S_j^* depends on σ and the hyperparameters, but does not depend on θ. The following result shows that for any fixed S_j^*, the posterior can be integrated over θ_j in closed form.

Proposition 9 *Let $P_{\mathbf{\Lambda}}(\sigma,\theta)$ be defined as in (16) and S_j^* be defined by (17). Then,*

$$P_{\mathbf{\Lambda}}(\theta_j\,|\,\sigma) \;=\; Beta_{S_j^*,\nu+1}(e^{-\theta_j}),$$

where $Beta_{\alpha,\beta}$ denotes the Beta distribution.

Proof sketch Replacing $\psi(\theta_j)$ with its value (5) yields

$$P_{\mathbf{\Lambda}}(\theta_j\,|\,\sigma) \propto e^{-S_j^*\theta_j}(1-e^{-\theta_j})^{\nu},$$

from which the desired result follows by a change of variable. □

As a consequence, we have that

$$P_{\mathbf{\Lambda}}(\sigma) \propto \prod_{j=1}^{t} Beta(S_j^*(\sigma), 1+\nu). \tag{18}$$

In the above, the notation $Beta(x,y)$ is used to denote the special function Beta defined as $Beta(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$.

We have shown thus that closed form integration over the continuous parameters θ_j is possible. The summation over the discrete parameters poses much harder problems. We list them here.

[5] The general form of a conjugate prior may include factors in σ and θ which do not depend on Λ. For simplicity of the exposition, we do not consider such a form here.

A first unsolved question is the range of the variables S_j^*. While the s_j variables in the infinite GM model are always integers ranging from 0 to infinity, the S_j^* variables can have non-integer values if ν or Λ_j are non-integer. The latter is almost always the case, since under the conditions of Proposition 8, Λ_j is not integer unless all its elements are 0 or 1. Second, Λ_j must have an infinite number of non-null entries, which may create problems for its numerical representation. And finally, there can be dependencies between S_j^* values for different j's. Hence, the factored expression (18) *should not be interpreted* as implying the independence of the S_j^*'s.

We illustrate these points by a simple example. Assume that the conjugate prior hyperparameters are equivalent to the fictitious sample $\{\pi_1 = (1|2|3|\ldots), \pi_2 = (2|1|3|\ldots)\}$. Then,

$$\nu = 2,\ \lambda_1 = \begin{bmatrix} 0.5 \\ 0.5 \\ 0 \end{bmatrix} \quad \Lambda_2 = \begin{bmatrix} - & 0.5 & 0 \\ 0.5 & - & 0 \\ 0 & 0 & - \end{bmatrix}, \quad \Lambda_3 = \begin{bmatrix} - & 0 & 0 \\ 0 & - & 0 \\ 1 & 1 & - \end{bmatrix},$$

$$R_1^0 = \begin{bmatrix} - & 0.5 & 0.5 \\ 0.5 & - & 0.5 \\ 0 & 0 & - \end{bmatrix}, \quad R_2^0 = \begin{bmatrix} - & 0 & 0.5 \\ 0 & - & 0.5 \\ 0 & 0 & - \end{bmatrix}, \quad R_3^0 = \begin{bmatrix} - & 0 & 0 \\ 0 & - & 0 \\ 0 & 0 & - \end{bmatrix}.$$

For this example, there are two central rankings $\sigma_1 = (1|2|3|\ldots)$ and $\sigma_2 = (2|1|3|\ldots)$ which have the same $S_{1:3}^* = (1, 0, 0)$, but no σ with $S_j^* \equiv 0$. Assume now that we are given just $R_{1:3}^0$, $\nu = 2$ and $S_1^*(\sigma) = 1$ for some σ. Because $S_1^* + 2$ is the sum of ranks of $\pi^{-1}{}_{1,2}(1)$ in σ, we can easily infer that the first two items in σ must be either $(1|2)$ or $(2|1)$ since any other σ will have $S_1^* \neq 1$. But, for either of these possibilities, the computation of $S_2^*(\sigma)$ from R_2^0 gives $S_2^*(\sigma) = 0$. Hence, knowing S_1^* informs about S_2^* (in fact determines it completely), showing that S_1^*, S_2^* are not independent.

Due to the above difficulties, computing the normalization constant of the posterior is an open problem. However, under some restrictive conditions, we are able to compute the normalization constant of the posterior in closed form.

Proposition 10 *If ν and $\Lambda_{1:t}$ are all integer, the S_j^* variables are independent, and the range of values of S_j^* is $\mathbb{P}$ then*

$$P(S_j^* = k) \ = \ (N+\nu) Beta(k+1, N+1+\nu),$$

and consequently

$$P_{\Lambda,\nu}(\theta_{1:t}, S_{1:t}^*) \ = \ (N+\nu)^t e^{-\sum_{j=1}^t [\theta_j S_j^* + (N+\nu)\ln\psi(\theta_j)]}.$$

The proof is given in Appendix A.1.

Now we examine the case of a single parameter IGM. The conjugate prior is given by $\nu > 0$ and a single matrix R^0 corresponding to the normalized sufficient statistics matrix R,

$$P_{\nu,R^0}(\theta,\sigma) \ \propto \ e^{-\theta L(\Sigma^T \nu R^0 \Sigma) + t\nu \ln\psi(\theta)}, \tag{19}$$

The posterior is

$$P_{\nu,R^0}(\theta,\sigma \,|\, \pi_{1:N}) \ \propto \ e^{-L(\theta\Sigma^T(\nu R^0 + R)\Sigma) + t(\nu+N)\ln\psi(\theta)]}.$$

Denote for simplicity $N' = N + \nu$, $R' = (R + \nu R^0)/(N+\nu)$, $S^*(\sigma) = L_\sigma(N'R')$. Then, the parameter θ again follows a Beta distribution given σ,

$$P_{N',R'}(\theta \,|\, \sigma) \ \propto \ Beta_{S^*; tN'+1}(e^{-\theta}).$$

After integrating θ out, we obtain

$$P_{N',R'}(\sigma) \propto Beta(S^*(\sigma), tN'+1).$$

Finally, let us note that that the priors in (16) and (19) are both informative with respect to σ. By replacing the term $L_\sigma(R_j^0)$ (respectively $L_\sigma(R^0)$) with some $r_j > 0$ (respectively $r > 0$) one can obtain a prior that is independent of σ, hence uninformative. However, this prior is improper.

6. Experiments

In this section, we conduct experiments on single θ estimation, general θ estimation, real data sets, and clustering.

6.1 Estimation Experiments, Single θ

In these experiments we generated data from an infinite GM model with constant $\theta_j = \ln 2, \ln 4$ and estimated the central permutation and the parameter θ. To illustrate the influence of t, t_π was constant over each data set. The results are summarized in Table 1.

Note that while θ appears to converge, the distance $d_K(\sigma^{ML}, \sigma)$ remains approximately the same. This is due to the fact that, as either N or t increase, n, the number of items to be ranked, increases. Thus the distance d_K will be computed between ever longer permutations. The least frequent items will have less support from the data and will be those misranked. We have confirmed this by computing the distance between the true σ and our estimate, restricted to the first t ranks. This was always 0, with the exception of $n = 200, \theta = 0.69, t = 2$ when it averaged 0.04 (2 cases in 50 runs) (A more detailed analysis of the ordering errors will be presented in the next subsection.)

Even so the table shows that most ordering errors are no larger than 1. We also note that the sufficient statistic R is an unbiased estimate of the expected R. Hence, for any fixed length $\tilde{t}$ of σ^{ML}, the σ estimated from R should converge to the true σ (see also Fligner and Verducci, 1988). The θ^{ML} based on the true σ is also unbiased and asymptotically normal.

6.2 Estimation Experiments, General θ

We now generated data from an Infinite GM model with $\theta_1 = \ln 2$ or $\ln 4$ and $\theta_j = 2^{-(j-1)/2}\theta_1$ for $j > 1$. As before, t_π was fixed in each experiment at the values $2, 4, 8$. We first look at the results for $t = 8$ in more detail. As the estimation algorithm has local optima, we initialized the θ parameters multiple times. The initial values were (i) the constant value 0.1 (chosen to be smaller than the correct values of all θ_j), (ii) the constant values 1 and respectively 2 depending whether $\theta_1 = \ln 2$ or $\theta_1 = \ln 4$ and (iii) the true θ parameters. The case (ii) ensured that the initial point is higher than all correct values for all the estimated θ_j.

Figure 6 shows the estimated values of θ_j for different sample sizes N ranging in $\{200, 500, 1000, 2000\}$. By comparing the respective (i) and (ii) panels, one sees that the final result was insensitive to the initial values and always close to the true θ_j. The results were also identical to the results for the initialization (iii), and this was true for all the experiments we performed. Therefore, in the subsequent plots, we only display results for one initialization, (i).

Qualitatively, the results are similar to those for single θ, with the main difference stemming from the fact that, with decreasing θ_j values, the sampling distribution of the data is spread more, especially w.r.t the lower ranks.

Estimates of θ (mean stdev)									
θ	**N**	200		500		1000		2000	
		mean	std	mean	std	mean	std	mean	std
0.69	$t=2$	0.68	0.04	0.68	0.03	0.68	0.03	0.68	0.024
	$t=4$	0.67	0.03	0.69	0.02	0.69	0.01	0.69	0.01
	$t=8$	0.68	0.02	0.69	0.01	0.69	0.01	0.69	0.007
1.38	$t=2$	1.34	0.13	1.37	0.09	1.39	0.05	1.37	0.04
	$t=4$	1.40	0.06	1.38	0.05	1.39	0.03	1.38	0.03
	$t=8$	1.37	0.03	1.38	0.03	1.38	0.02	1.38	0.01

Ordering error									
θ	**N**	200		500		1000		2000	
		$d_K=0$	$d_K=1$	$d_K=0$	$d_K=1$	$d_K=0$	$d_K=1$	$d_K=0$	$d_K=1$
0.69	$t=2$	0.42	0.36	0.28		0.36		0.40	0.38
	$t=4$	0.36	0.36	0.40		.44		0.32	0.30
	$t=8$	0.32	0.34	0.44		0.40		0.38	0.32
1.38	$t=2$	0.82	0.18	0.92	0.08	0.76	0.24	0.90	0.08
	$t=4$	0.92	0.08	0.92	0.08	0.88	0.12	0.88	0.10
	$t=8$	0.84	0.16	0.92	0.08	0.76	0.20	0.16	0.88

Number observed items *n*									
θ	**N**	200		500		1000		2000	
		mean	std	mean	std	mean	std	mean	std
0.69	$t=2$	9.24	0.92	10.76	0.66	11.88	1.16	12.68	1.07
	$t=4$	11.92	0.81	12.92	1.04	14.32	1.10	14.88	1.01
	$t=8$	16.04	0.89	17.36	0.99	18.16	1.07	19.40	0.91
1.38	$t=2$	5.68	0.74	6.04	0.79	6.76	0.78	7.32	0.55
	$t=4$	7.72	0.84	8.16	0.74	8.68	0.75	9.48	0.82
	$t=8$	11.52	0.65	12.52	0.58	13.24	0.66	13.40	0.71

Table 1: Results of estimation experiments, single parameter IGM. Top: mean and standard deviation of θ^{ML} for two values of the true θ and for different t values and sample sizes N. Middle: the proportion of cases when the ordering error, that is, the number inversions w.r.t the true σ^{-1} was 0, respectively 1. Bottom: number of observed items n (mean and standard deviation). Each estimation was replicated 25 or more times.

This figure allows us to observe the "asymmetry" of the error in θ^{ML}. The estimates seem to biased towards larger values, especially for higher j and less data. There is a theoretical reason for this. Recall that by Equation (12) θ is a decreasing function of $L_\sigma(R)$. If the true σ is not optimal for the given R, due to sample variance, then θ^{ML} will tend to overestimate θ. Hence θ^{ML} is a *biased* estimate of θ. If however, due to imperfect optimization, the estimated σ^{ML} is not optimal and has higher cost than σ, then θ^{ML} will err towards underestimation. In Figures 6 and 7 the bias is always positive, indicating that the minimization over σ is done well (even though it is not guaranteed to reach optimality).

Now we consider the recovery of the central permutation σ. From our previous remarks, we expect to see more ordering errors in the bottom ranks of σ^{ML}, where the distribution is less concentrated (smaller θ_j) and there is less data available. To visualize these effects easier, it is interesting to look at rankings with small t. When t is small, 2, 4 or 8, the total number of items to be ranked (Figure 8) is several times larger than t for our experiments. Thus the estimation algorithm has to put together an ordering over this many items, when only groups of $t = 2, 4, \ldots$ were observed together. As an additional confounding effect, the top elements will be oversampled, so the information about the lower ranks will have to be inferred indirectly by pooling the information from the whole data set. This is what the algorithm is doing, and Figure 9 shows how well it succeeds in that.

The figure displays the ordering error between σ^{ML} and the true σ for each rank, which is measured by $s_j(\sigma^{ML}|\sigma)$. Recall that the total number of inversions between σ^{ML} and σ is the sum of all s_j; similarly, the total number of errors in σ^{ML} up to rank r is given by $\sum_{j=1}^{r} s_j(\sigma^{ML}|\sigma)$. All plots illustrate the same general tendency of the s_j values to increase *slowly* with j. The increase is slower when there are more observations per rank, that is when N and t are larger, and when the θ_j are larger (thus the data distribution is more concentrated).

6.3 Experiments on Real Data Sets, With General θ, Tied Parameters

The next experiment was conducted with the data collected by Cohen et al. (1999). The data consists of a list of 157 universities, the queries, and a set of 21 search engines, the "experts". Each search engine outputs a list of up to $t_{max} = 30$ URL's when queried with the name of the university. The data set provides also a "target" output for each query, which is the university's home page.

Hence, we have 147 estimation problems (10 universities with no data), with sample size $N \leq 21$ (as some experts return empty lists) and with variable length data ranging from $t = 1$ to $t = 30$. Figure 10 gives a summary view of number of samples for each rank $N_j,\ j = 1:t$, the number of distinct items n and the cumulative number of ranks observed (i.e., $T = \sum_{j \leq 30} N_j$). These values suggest that estimating a fully parameterized model with distinct $\theta_{1:30}$ may lead to overfitting and therefore we estimate several parameterizations, all having the form $\Theta_r = (\theta_1, \theta_2, \ldots \theta_{r-1}, \theta_r, \theta_r, \ldots \theta_r)$. In other words, ranks $1 : r-1$ have distinct parameters, while the following ranks share parameter θ_r. We call $\theta_{1:r-1}$ the *free* parameters and θ_r the *tied* parameter. For $r = 1$ we have the single parameter model, and for $r = t_{max} = 30$ we have the fully parameterized model.

Estimating a model with r parameters is done by a simple modification of the ESTIMATESIGMATHETA algorithm which is left to the reader.

The estimation algorithm was started from the fixed value $\theta_j = 0.1$ for all runs. The number of iterations to convergence range between 10 and 50, with typical value 18. The running time was around per model estimated.

In Figure 11 we give a synopsis of the values of the θ parameters under different models. The single parameter models yields θ values in the range [0.007, 0.104] with the 10%, 50% and 90% quantiles being respectively 0.009, 0.018,and 0.032. The parameters θ are on average decreasing in all models, with the free parameters higher than the tied parameters for the remaining ranks. This is true on average only, while for individual samples some of the free parameters may be smaller than the tied parameter.

Notice also that for the models with fewer parameters the values of the free parameters tend to be higher than the corresponding values in models with more parameters. Compare for instance the values of θ_1 in the two-parameter model with θ_1 in the 30 parameter model.

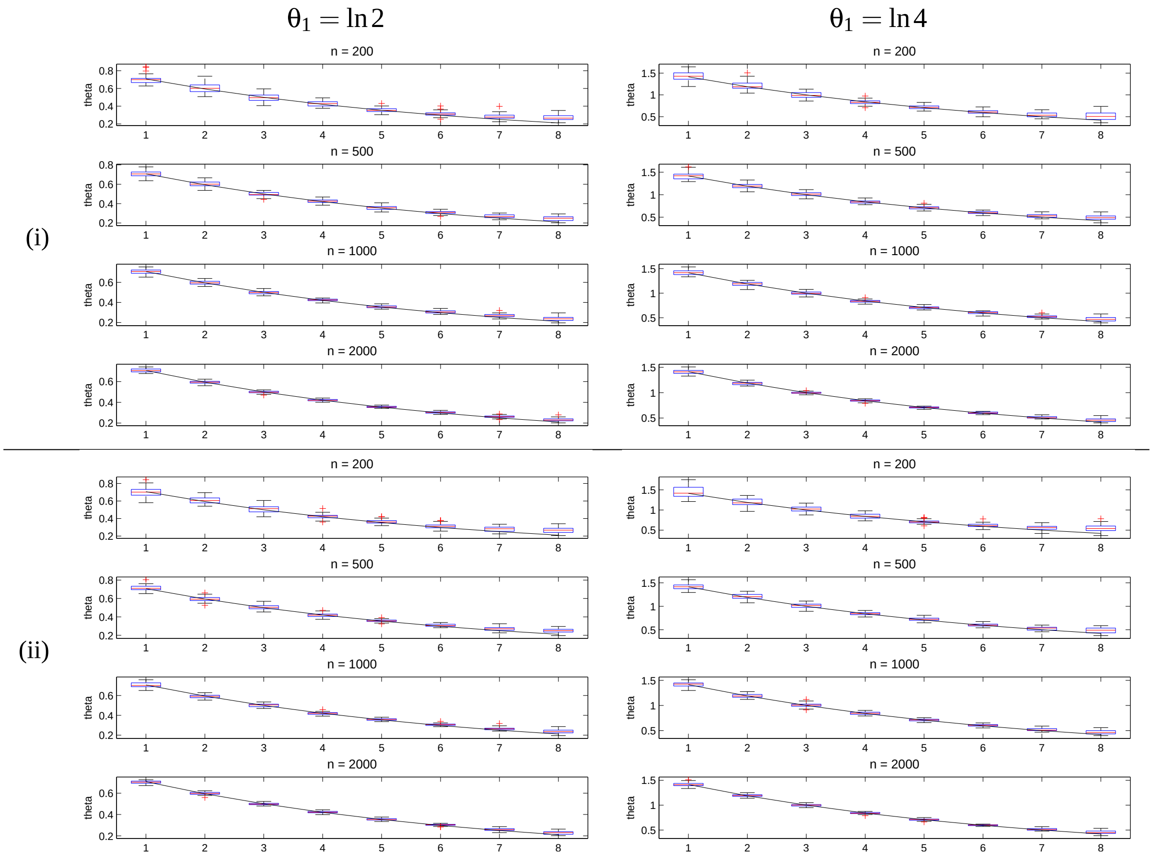


Figure 6: Estimation of the parameters $\theta_{1:t}$ for $t = 8$, different sample sizes $N = 200, 500, 1000, 2000$, different true parameters θ, and different initializations: (i) $\theta_j \leftarrow 0.1$, (ii) $\theta_j \leftarrow 1$ or 2. For each experimental condition, the corresponding graph displays box plots of the obtained estimates of θ_j, $j = 1:8$ for 50 random samples, with the j on the horizontal axis. The continuous line crossing the box plots marks the true values of the parameters $\theta_{1:8}$ (exponential decay starting from the given θ_1).

For each query and each model size, we computed the rank of the true university home page, that is, the *target*, under the estimated central permutation σ^{ML}. Assuming the search engines are reasonably good, this rank is an indirect indicator of the goodness of a model. In addition, for each query, we selected one model by BIC and calculated the target ranks for these models. Table 2 gives the mean and median of the target rank for each model, as well as for the BIC selection. The rank is assigned to $t_{max} + 1 = 31$ if the target is not among the items returned by the search engines.

It is evident that while BIC's performance is better than selecting a one-parameter model, it is not optimal w.r.t the ranking of the target home page.

We used a modified form of the BIC criterion, that takes into account that the continuous parameters are not all estimated from the same sample size. We have derived[6] the following formula

[6] The derivation is omitted, being outside the scope of this paper.

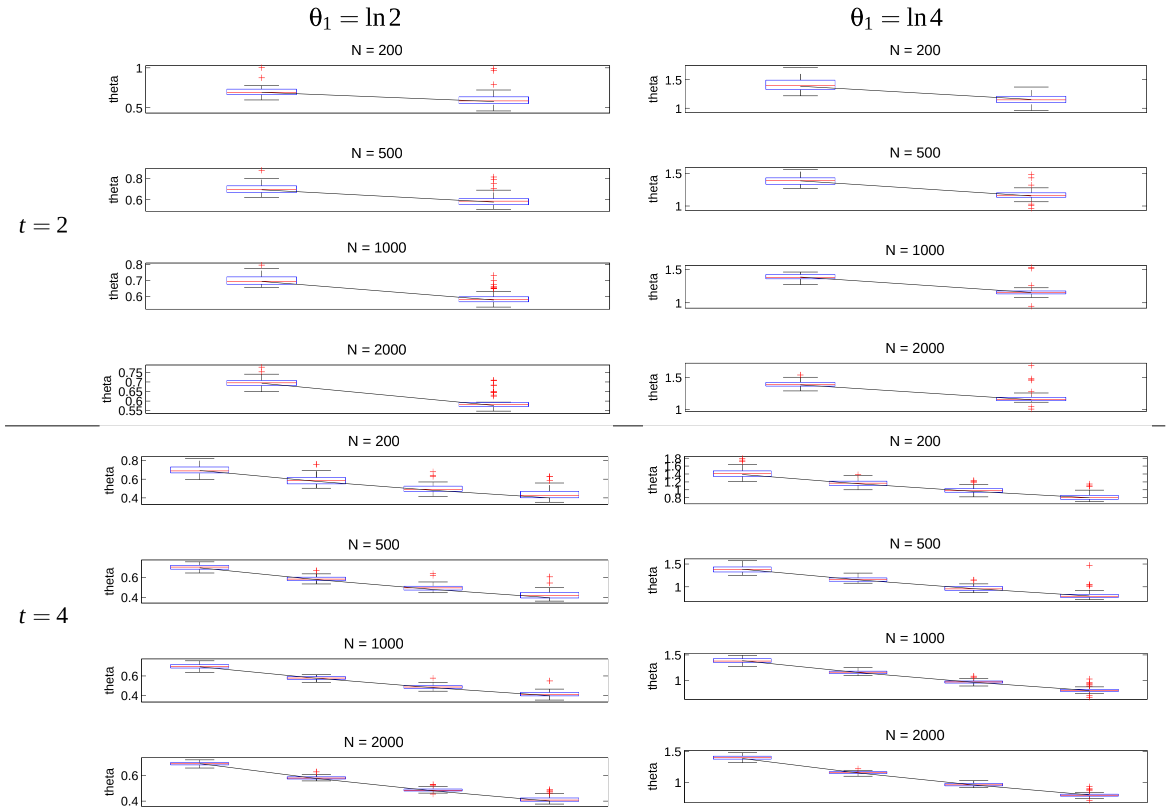


Figure 7: Same as Figure 6 for $t = 2$ and $t = 4$ and a single initial point $\theta_j \equiv 0.1$.

for BIC:

$$BIC(r) \; = \; \ln P_{\sigma,\theta}(\mathcal{S}_N) - \frac{1}{2}\sum_{j=1}^{r} \ln N'_j \tag{20}$$

where $N'_j = N_j$ for $j < r$ and $N'_r = \sum_{j'=r}^{t} N_j$. This expression approximates the marginal distribution of the model w.r.t the continuous parameter. The discrete parameter σ is not marginalized out. This parameter has always the same dimension, dictated by the observed data, and independent of r. In any finite data situation, the parameter will be finite. Therefore we can see the model selection problem as a model selection over r and a very large but finite set of discrete σ's. Maximizing the BIC in (20) is equivalent with maximizing the BIC over this much larger set of models, if one assumes that the ML estimation procedure attains a global optimum.

Next, we tested the ESTIMATESIGMATHETA algorithm on the Jester data of Goldberg et al. (2001).[7] This data set represents a set of 100 jokes, which were scored by approximately 25,000 people. From the numerical scores, we obtained a partial ordering over the jokes rated by each individual. Mao and Lebanon (2008) also analyzed this data set and found that it was multimodal. To obtain data sets closer to unimodality, we picked a person at random (this is person 945 in the data) and extracted the N nearest neighbors of this ranking, for $N = 200$ and $N = 12,000$. The smaller data set was expected to be more concentrated than the larger data set.

[7] Available at `http://goldberg.berkeley.edu`.

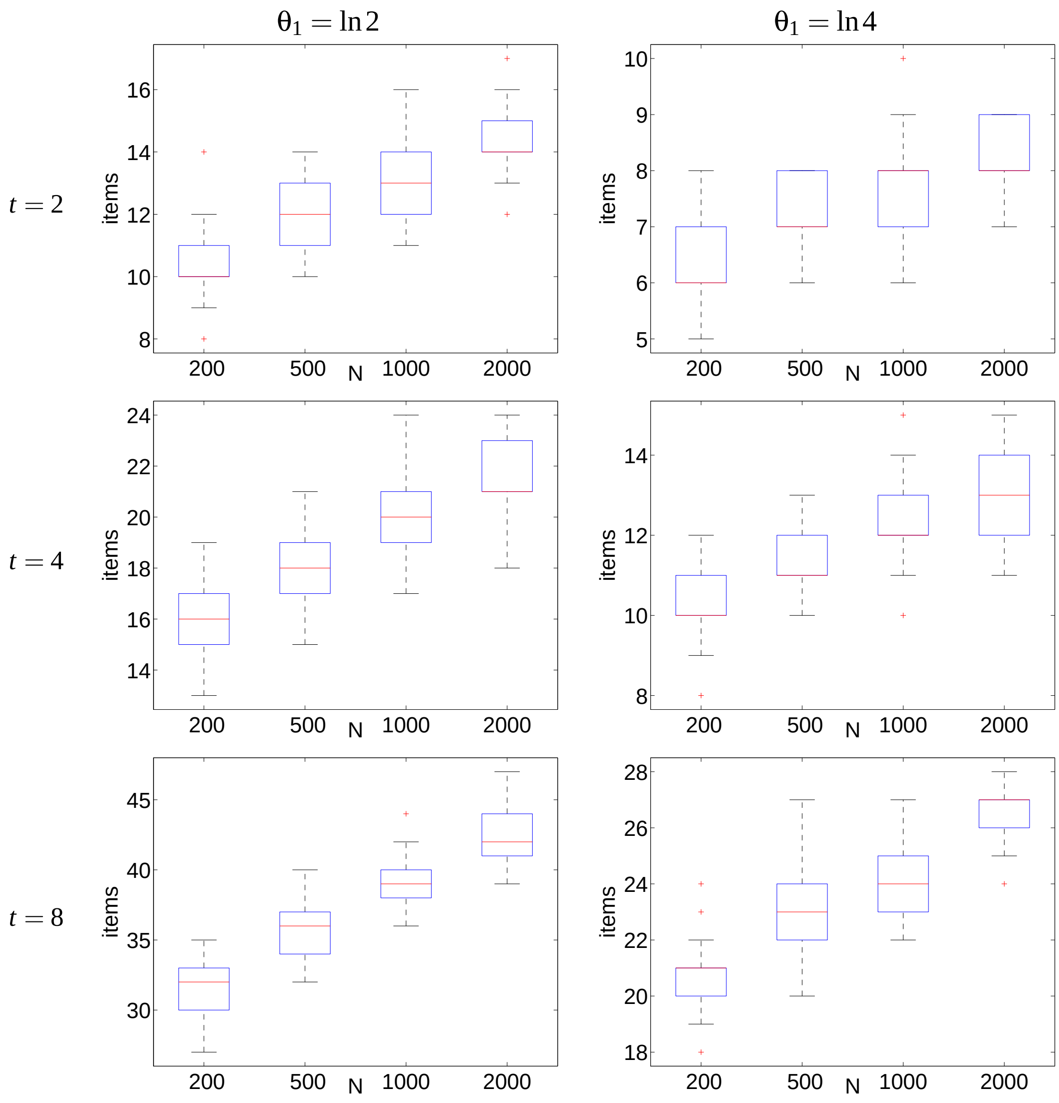


Figure 8: Number of items observed for $N = 200, 500, 1000, 2000$, different parameters θ and $t = 2, 4, 8$. Each box plot represents the distribution of the number of items over 50 random samples.

We ran the ESTIMATESIGMATHETA on this data set with different values of r. The log-likelihoods obtained on the training set and the BIC values are shown in Figure 12.

As expected, the likelihood is highest or nearly so for the model with maximum number of free parameters (for $N = 12,000$ the likelihood is not monotonic due to imperfect optimization over σ). However, the BIC is not monotonic. For the large N case, where the data is dispersed, it chooses a model with $r = 76$ parameters. The estimated θ_j values range in $[0.05, 0.07]$ for $j = 1:30$ (nearly all ranks that have $N_j = 12000$) but become higher, up to 0.2 for the ranks that have smaller N_j's. For the smaller and more concentrated data set, BIC has equal values for three different models: the

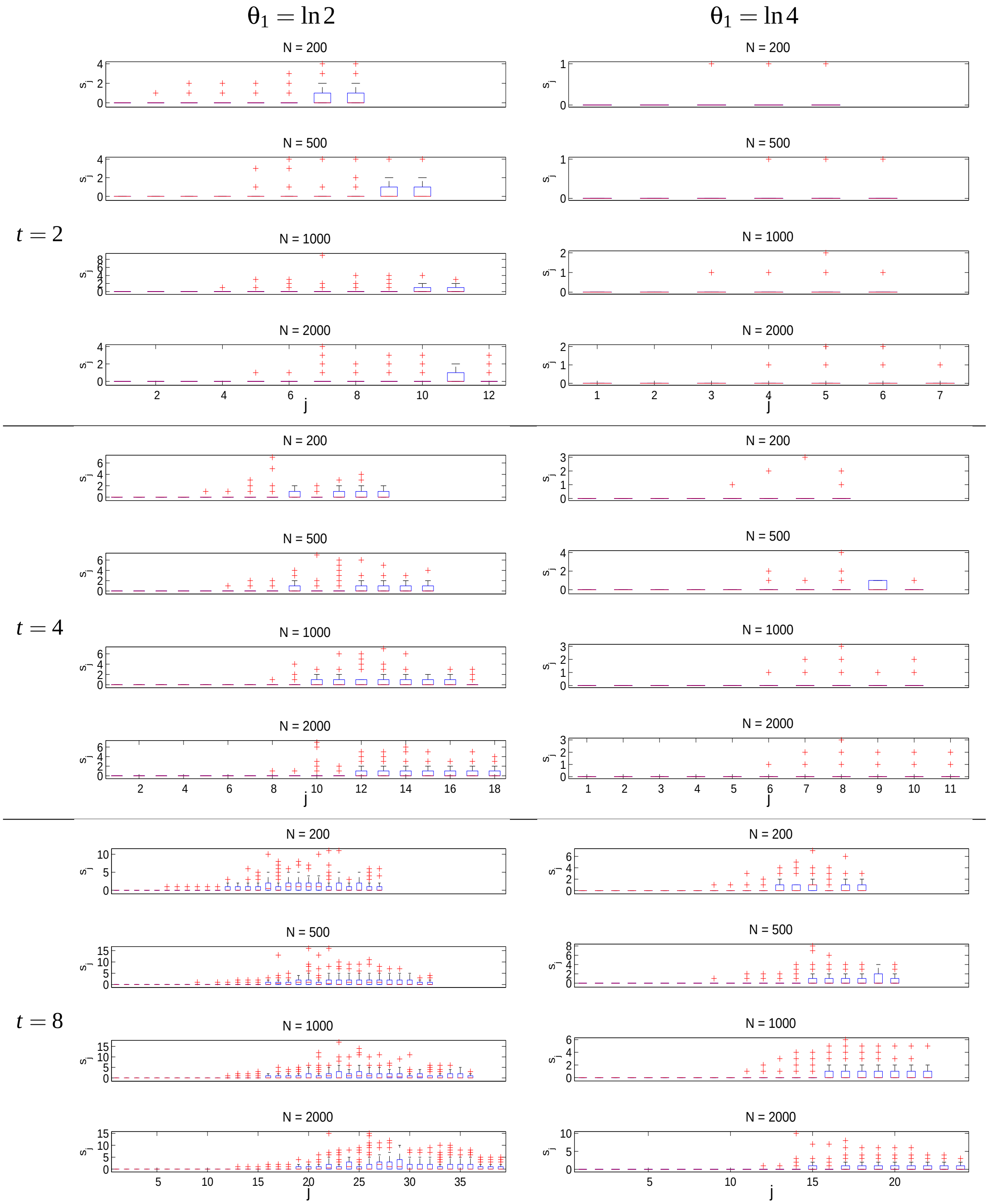


Figure 9: Ordering errors for $N = 200, 500, 1000, 2000$, different parameters θ_j , and $t = 2, 4, 8$. The error for a rank j is given by $s_j(\sigma^{ML}|\sigma^{true})$. Each box plot represents the distribution of s_j over 50 random samples.

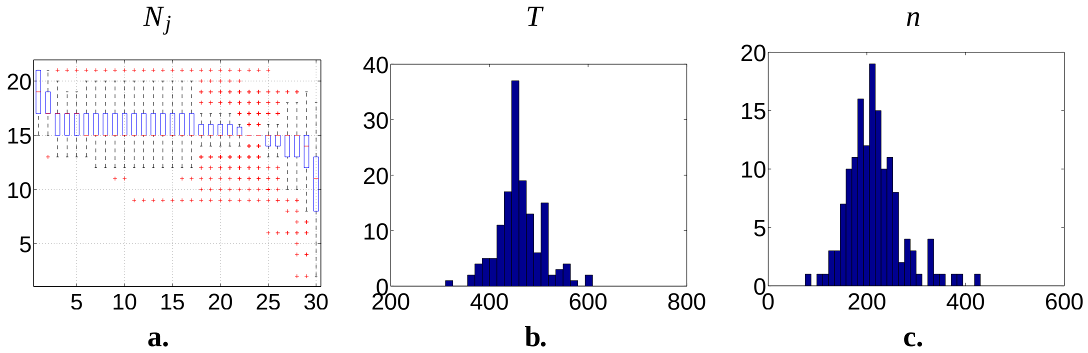


Figure 10: Summaries of the universities data: boxplots of the number of samples per rank, (a), histogram of total items observed T (b), histogram of the number n of distinct items observed (c), over all 147 queries.

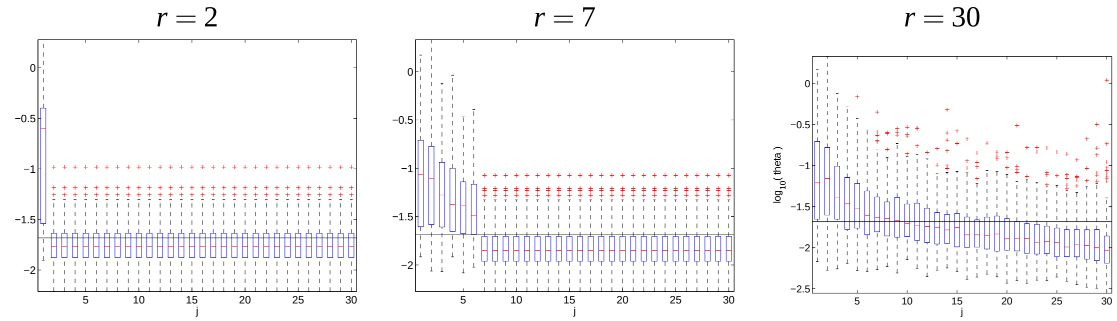
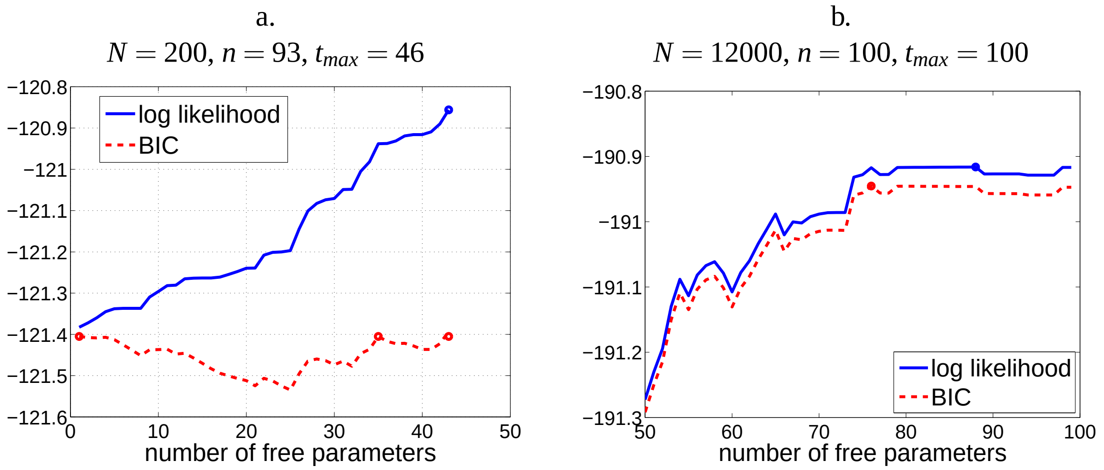


Figure 11: Boxplots of the Θ estimates over all queries for model with 2, 7, and 30 parameters. The vertical axis of the scale is *logarithmic, base 10*, that is, 0 corresponds to $\theta_j = 1$ and -2 to $\theta_j = 0.01$. For clarity, the distribution of the tied parameter (which is always the last parameter) is replicated for $j = r : t_{max}$. The horizontal line marks the mean value of θ in the single parameter model.

Model size	1	2	3	4	5	6	7	8	9	10	30	BIC
Mean rank (good)	5.3	5.7	4.2	4.2	**4.1**	**4.1**	4.4	4.5	5.0	5.2	5.1	5.5
Median rank (good)	3	3	**1.5**	2	2	2	2	2	2	3	3	3
Mean rank (all)	16.5	16.1	**15.4**	15.5	15.5	15.6	15.8	15.7	15.9	16.0	16.0	17.5
Median rank (all)	13	15	11	11	12	**9**	10	10	11	11	11	18

Table 2: Mean and median of the rank of the target web page under each model, and under the BIC selected model. These statistics are computed once over all 147 universities and once over a subset of 41 universities where the target is always ranked in the first 30; the subset is labeled as “good”.


a.

$N = 200$, $n = 93$, $t_{max} = 46$

b.

$N = 12000$, $n = 100$, $t_{max} = 100$

runtime 14.5–15s/model, $\theta_{1:30} \approx 0.1$
Mallows' $\theta = 0.1$

runtime 88–90.5s/model, $\theta_{1:30} \approx 0.05 - 0.07$
Mallows' $\theta = 0.076$, log-likelihood, BIC=-194.1

Figure 12: Log likelihood and BIC values per data point on two subsamples from the Jester data set. The circles mark the maxima of the log-likelihood, respectively BIC.

single parameter model, the full parameter model, and an intermediate model with $r = 36$. This is not so suprising as it may appear, since $N_j = N$ for $j \geq 36$, and the estimated θ_j values in this range are all very close to 0.1. Thus, the three models will effectively differ only in the way they model ranks $37 : 43$.

This finding suggests that the single parameter (infinite) Mallows model is a good model locally, in the space of the `jester` data. The large N experiment indicates otherwise, which agrees both with our common sense assumption that this data is multimodal, and with the finding of Mao and Lebanon (2008). In Mao and Lebanon (2008) a non-parametric Mallows model was used (more about this model in Section 7.3); our experiment supports their use of a single parameter model.

6.4 Clustering Experiments

The first experiment was with artificial data. We generated sample orderings with 3 clusters of 150 rankings each. Each cluster k was sampled from an Infinite GM model with a single spread parameters θ_k, with θ_1, θ_2, θ_3 equal to 1.5, 1.0, 0.7 respectively. The cluster centers are random permutations of infinitely many objects. In addition, each data set contains 50 outliers. In each data set all data had the same length t. We experimented with $t_\pi = 4, 6, 8$.

We ran the Exponential Blurring Mean-Shift, K-means, and EM Model-based clustering algorithms 10 times on samples from this distribution. For EBMS, the scale parameter was estimated based on the average of pairwise distances. In step 5c of the algorithm, the new ranking can be much longer than the original partial ranking. As seen above, the last ranks are subject to noise and overfitting. Therefore we truncated the new ranking to the length of number of observed items.

For the K-means and model-based algorithms, we experiment with different numbers of clusters, and report the best classification error with respect to the true clustering. This puts these two algorithms at an advantage w.r.t EBMS, but as Table 3 shows, even so the nonparametric algorithm achieves the best performance.

Note that the error rate in Table 3 is computed including the outliers, that is, we compared a true clustering with 53 clusters (3 clusters and 50 singletons) to the clustering obtained when the algorithm converged.

For EM and K-means the number of clusters associated with the lowest classification errors was between 3 and 5. From the table we see that the K-means and model based approach identified three primary clusters correctly. K-means did not have the ability of identifying the outliers, so it just assigned each outlier into one of those primary clusters. The model-based approach assigned outliers into primary clusters too, but it also gave more uncertainty on the outliers (the probabilities of outliers belonging to their assigned cluster were relatively smaller than data from primary clusters).

The running time per data set of EBMS was under a minute, and the number of iterations to convergence followed the pattern typical of mean-shift algorithms and was never larger than 10.

Next, we examine data on college course preferences in the Republic of Ireland. Each prospective student applies by ranking up to 10 degree course in order of preference. Extensive details of the college applications system are available at `http://www.cao.ie`.

The data used in our analysis was previously studied by Gormley and Murphy (2006). They found that the geographical positions of the institution had a significant influence on choice of courses which complicated the interpretation of the vocational callings. Our analysis focuses on the subset of students who applied to Trinity College Dublin (TR) and University College Dublin (DN), both located in the capital of Ireland. These two universities offered $n = 228$ degree courses. There were 1095 female and 862 male applicants who only put TR and DN courses in their top-5 preferences. The EBMS algorithm is applied to top-5 rankings for the female and male applicants separately.

Table 4 shows these clustering results. For the female applicants the first cluster mostly consists of Art, Law and Business courses. Since the largest cluster contains about three quarters of the data, we run the EBMS clustering again for the applicants within this cluster and find four major subgroups in term of vocational callings: Law, Business, Drama, and English. The clustering results for male applicants show 5 clusters, plus a singleton. We run EBMS again for the largest cluster, and find three major sub-groups in term of vocational callings: Finance, Law, and History. As Gormley and Murphy (2006) discovered in their experiments, for the subset of Dublin applicants, there are differences between the clustering of females and males in the central permutations, but similarities too. The main similarity is that the grouping is vocational. Each group contains courses in both universities, with no strong preference for one versus the other. Second, a large proportion of both genders opt for business, economics and law disciplines. For the females, the Arts courses are also highly favored. As Gormley and Murphy (2006) explains, the Arts course is a broad liberal arts

Top-t rankings	EBMS	K-means	EM
$t = 4$	0.0030 (0.0001)	0.1014 (0.0038)	0.1008 (0.0025)
$t = 6$	0.0014 (0.0001)	0.0986 (0.0010)	0.1000 (0.0000)
$t = 8$	0.0002 (0.0001)	0.0972 (0.0010)	0.1000 (0.0000)

Table 3: Classification Errors: mean and standard deviation of 10 random samples.

course which can be followed with many different specializations later on. Hence, its high ranking in several clusters is an indication that female candidates want to leave more options open later. In general, the central rankings of each cluster are very clearly separating the pool of candidates into various profiles.

Males

cluster size	657	143	35	13	13
1st choice	BESS (TR)	Engineering (TR)	Science (DN)	Science (DN)	Medicine (TR)
2nd choice	Commerce (DN)	Science (DN)	Science (TR)	Science (TR)	Medicine (DN)
3rd choice	Business and Law (DN)	Engineering (DN)	Mathematics (TR)	Medicine (TR)	Pharmacy (TR)
4th choice	Arts (DN)	Computer Science (DN)	Theoretical Physics (TR)	Medicine (DN)	Dental Science (TR)
5th choice	Economics and Finance (DN)	Computer Science (TR)	Theoretical Physics (DN)	Pharmacy (TR)	Veterinary Medicine (DN)

Females

cluster size	725	162	141	41	26
1st choice	Arts (DN)	Arts (DN)	Arts (DN)	Physiotherapy (DN)	Physiotherapy (TR)
2nd choice	Law (DN)	Psychology (DN)	Psychology (DN)	Physiotherapy (TR)	Physiotherapy (DN)
3rd choice	BESS (TR)	Psychology (TR)	Psychology (TR)	Radiation Therapy (TR)	Science (DN)
4th choice	Business and Law (DN)	Science (DN)	Law (DN)	Radiography (DN)	Medicine (DN)
5th choice	Law (TR)	Science (TR)	Social Science (DN)	Occupational Therapy (TR)	Medicine (TR)

Table 4: EBMS Clustering of female and male applicants. BESS stands for Business, Economic and Social Science, TR for Trinity College Dublin and DN for University College Dublin.

7. Discussion and Related Work

In this section, we discuss the relation between IGM and GM, other related models and algorithms, and draw the brief conclusion.

7.1 Relation to the GM Model

It is useful to compare the various aspects of the IGM presented here with the respective aspects of the standard GM. We do so now, highlighting also which of them were already published and which are new.

- *s* **representation.** This was introduced by Fligner and Verducci (1986) for the GM model. For finite number of items n, j ranges in $1:n-1$ and s_j in $0:n-j+1$.

- **marginal distributions,** $P_{\theta,\sigma}$**, over top-**t **orderings.** It is also introduced by Fligner and Verducci (1986). The main difference with the IGM model is in the normalization constant, which has the expression

  $$\prod_{j=1}^{t} \psi_j(\theta_j) \quad \text{with} \quad \psi_j(\theta_j) \;=\; \frac{1-e^{-(n-j+1)\theta_j}}{1-e^{-\theta_j}}. \tag{21}$$

  Also, for the GM model, the underlying space $\mathcal{S}_n$ is finite, while for the IGM, $\mathcal{S}_{\mathbb{P}}$ is uncountable.

- **sufficient statistics as in Proposition 1.** Proposition 1 represents a new result for the GM as well. The only difference is in the replacement of $\psi(\theta_j)$ with $\psi_j(\theta_j)$ from (21) in (8). The nearest previous result is that of Meilă et al. (2007) which establishes sufficient statistics for the GM model over complete permutations.

If we have a single parameter GM model and complete permutations, then it is easy to see that $\sum_j Q_j$ represents the sufficient statistics for σ alone. In computer science estimating σ in this context is called the *consensus ranking* problem, or the *minimum feedback arc set* problem. In this case, by setting $t = n$ for all permutations, the sufficient statistics defined in (8) reduce to the previously known $\sum_j Q_j$. Thus, our main contribution in this respect is to prove that not knowing θ, and not observing complete permutations still results in an exponential family model with sufficient statistics.

- **θ estimation.** In the GM case, this is a convex unidimensional optimization solved numerically (Fligner and Verducci, 1986; Meilă et al., 2007).

- **σ estimation by** SIGMA* The SIGMA* and ESTIMATESIGMATHETA algorithms can be used for the GM model as well, with the estimation of θ performed numerically. The closed form Equation (12) can serve as a very good initial point for the iterative optimization algorithm. Note that while the heuristics SORTR and GREEDYR are simple, they could not have been applied before to the GM model over top-t orderings because it was not known that this model has sufficient statistics for σ.

- **conjugate prior** Fligner and Verducci (1990) introduced an informative prior for θ, which had a single “sufficient statistics” parameter. They used it with a uniform prior over σ, noting that this prior cannot be normalized or integrated analytically. The informative conjugate prior for θ and σ introduced in Section 5 applies also to the standard GM. Again, the main formal change is replacing $\psi(\theta_j)$ by $\psi_j(\theta_j)$. With this change, we lose the elegant closed form integration over θ proved in Proposition 9. The GM conjugate prior will not be in general integrable in closed form over θ. The uninformative prior for the IGM becomes of course a proper prior in the GM case. If we set all the r_j parameters to the same value, we obtain exactly the same prior as Fligner and Verducci (1990).

  Often a non-informative prior is used as a regularizer for the ML estimator, turning it into a *Maximum A-Posteriori (MAP)* estimator. This is possible for the IGM and GM too. All one needs to do is to replace, in the inputs to the estimation algorithms, the data sufficient statistics with the posterior sufficient statistics.

- EBMS **clustering** The algorithm adapts seamlessly to finite number of items.

7.2 Other Models and Algorithms for Finite Permutations

This work acknowledges its roots in the work of Fligner and Verducci (1986) on stagewise ordering models and in the recent paper Meilă et al. (2007). The latter shows for the first time that GM models have sufficient statistics, and describes an exact but non-polynomial algorithm to find the central permutation. While similarities exist between the algorithm of Meilă et al. (2007) and the SIGMA* algorithm presented here, we stress that our representation (based on the inversion table $(s_j)_{j=1:t}$) is *different* from the representation (denoted V_j) in Meilă et al. (2007).

In fact, the two representations could be called reciprocal, as for any given complete permutation π, finite or infinite, $s_j(\pi|\mathrm{id}) = V_j(\pi^{-1}|\mathrm{id})$. This difference is trivial if complete permutations are observed, but not for missing data. In particular, the distribution of V_j for top-t orderings does not seem to have sufficient statistics for $j > 2$ even in the case of finite permutations. The s_j representation has another advantage that V_j has not: for any finite data set, a parameter s_j is either completely determined or completely undetermined the data, whereas in the reciprocal V_j representation *all* V_j are weakly constrained by data.

While both our SIGMA* and the algorithm of Meilă et al. (2007) perform branch-and-bound search on a matrix of sufficient statistics, the sufficient statistics in this paper are derived by an entirely different method, and cannot be obtained by naively replacing the sufficient statistics of Meilă et al. (2007).

It may be noted that the cost $L_\sigma(R)$ bears a striking similarity to the cost function used by Quadrianto et al. (2010) in the context of matching by the method of kernelized sorting. The latter cost function can be expressed as $\operatorname{trace} K\Sigma^T L\Sigma$, to be minimized over Σ. The authors show that this problem is quadratic in the matrix Σ when K, L are symmetric, positive definite matrices and use a quadratic relaxation algorithm to optimize it efficiently. The cost $L_\sigma(R)$ can be rewritten as $\operatorname{trace} Q_0\Sigma^T R\Sigma$ where Q_0 is the upper triangular matrix defined in the proof of Proposition 2. The main difference between our cost function and the one of Quadrianto et al. (2010) is the fact that ours involves the non-symmetric matrices Q_0, R and is not a quadratic problem in Σ.

An interesting application of the GM model to multimodal data is Lebanon and Lafferty (2003), where the σ's play the role of the data, so the parameter estimation is done entirely differently. In an early work Critchlow (1985) examines several classes of (Haussdorf) distances for partial orderings. Murphy and Martin (2003) cluster ranking data by the EM algorithm and in Gormley and Murphy (2005, 2006) the EM is used for the purpose of analyzing Irish voting patterns and college applications. The base model used by the latter papers is is not the Mallows model but the Plackett-Luce model (Plackett, 1975; Luce, 1959). The estimation of this model from data is much more difficult and, as Gormley and Murphy (2005) show, can be only done approximately. Busse et al. (2007) use the s_j representation in the context of EM clustering of partial orderings, without however recognizing the existence of sufficient statistics.

A greedy algorithm for consensus ordering with partially observed data is introduced in Cohen et al. (1999). Meilă et al. (2007) show that their cost function optimized is closely related to the log-likelihood of the Mallows' model, using a modified form of the Q matrix defined in (10). This algorithm, like GREEDYR, does not estimate a θ parameter. Cohen et al. (1999) introduce a computational improvement based on interpreting a value $Q_{ii'} > 0.5$ as an arc from i to i'. They note that it is sufficient to search for the optimal permutation in each *strongly connected component* of the resulting directed graph, which can sometimes greatly reduce the dimension of the search space.

If a permutation π is not complete, Cohen et al. (1999) replaces the unobserved $Q_{ii'}(\pi)$ with the value 0.5. This ad-hoc procedure allows the GREEDYR to run on top-t rankings, but it is not statistically correct, since the optimized cost will not be a likelihood. If we use the correct matrix of sufficient statistics R, then the reduction procedure based on strongly connected components does not apply any more.

7.3 Other Models and Algorithms For Infinite Permutations

All the above works deal with permutations on finite sets. In fairness to Cohen et al. (1999) we remark that their work, although non-rigorous with respect to incomplete permutations, is motivated by the same problem as ours, that is, dealing with a very large set of items, of which only some are ranked by the "voters".

The paper of Thoma (1964) studies the space of infinite permutations which differ from the identity in a finite number of positions. In the vocabulary of the present paper, these would be the infinite permutations at finite distance d_K from σ. In a single parameter infinite GM, these infinite permutations are the only ones which have non-zero probability. While from a probabilistic perspective the two views are equivalent, from a practical perspective they are not. We prefer to consider in our sample space all possibile orderings, including those with vanishing probability. It is the latter who are more representative of real experiments. For instance, in the university web sites ranking experiment, our model assumed that there is a "true" central permutation from which the observations were generated as random perturbations. This is already an idealization. But we also have the liberty to assume that the observations are very long orderings which are close to the central permutation only in their highest ranks, and which can diverge arbitrarily far from it in the latter ranks. We consider this a more faithful scenario than assuming in addition that the observation must be identical to the central permutation (and hence to each other!) on all but a finite number of ranks.

Recently Mao and Lebanon (2008) introduced a kernel density estimator and estimation algorithm, that elegantly allows partial orderings of a large variety of *types* to be modeled together. The kernel is the single parameter Mallows' model, with θ as kernel width. One of the interesting contributions of this paper is an algorithm for averaging the $d(\tilde{\pi}_1, \tilde{\pi}_2)$ over all (infinite) permutations $\tilde{\pi}_1, \tilde{\pi}_2$ compatible with given partial orderings π_1, π_2. The relation with EBMS is evident. It is also evident that within EBMS one could incorporate the average distance as calculated by Mao and Lebanon (2008) instead of the current set distance, with everything else staying the same. Since the average distance is always larger than the set (minimum) distance for top-t permutations, and in particular it is not 0, the optimal kernel width θ will have different values.

7.4 Conclusion

We have introduced a natural extension of stagewise ordering to the the case of infinitely many items. The new probabilistic model preserves the elegant properties of its finite counterpart: it has sufficient statistics, an exact estimation algorithm (albeit intractable in the worst case) and tractable heuristics that work well when the data come from a unimodal distribution. Sampling, distance computations, clustering extend to this class of models in a natural way.

We have paid particular attention to non-parametric clustering by mean-shift blurring, showing by experiments that the algorithm is practical and effective. This illustrates our view of the utility of the IGM model. The IGM, an exponential model with a simple, intuitive distribution, should be seen as a building block for more complex distributions, as needed by the data at hand. For instance, the extension to finite mixtures (that is, parametric clustering and multimodal distributions) is immediate. It is also an open problem to extend the kernel density estimator of Mao and Lebanon (2008) to infinite models and GM models with multiple parameters.

We are not aware of any statistical work on estimating parametric models over infinite orderings. There are also no previous results on sufficient statistics for finite partial orderings, so the present paper can be said be first in this respect as well.

One important advantage of having a model with multiple parameters, with n finite or infinite, is that each rank can be modeled by a separate parameter θ_j (the standard GM/IGM) or one can tie the parameters of different ranks (the way we did in Section 6). This way, one can use larger θ_j values to penalize the errors in the first ranks more, and smaller θ_j values for the lower ranks, where presumably more noise in the orderings is permissible. This property of the model fits well with human perception of “distance” between orderings, or with the noise we may expect in human generated data.

Tying the parameters for the lower ranks is also important for reducing variance. If the observed data have different lengths t, then necessarily $N_j \geq N_{j'}$ for $j' > j$. In other words, for larger j’s we may have less data available to estimate θ_j. Tying the parameters has the benefic effect of smoothing the θ_j values, as it was shown in all the experiments with real data. Another way to smooth the parameters is to use an uninformative prior as a regularizer. This way, too, each θ_j can be regularized separately by the hyperparameter r_j. This hyperparameter has a clear meaning—it is the expectation of s_j in the fictitious sample; therefore, a user can easily tune the strength of the prior using Equation (7) with r_j in place of ξ_j. This equation will give for any r_j a value θ_j^0 towards which the θ_j value will be shrunk. Shrinkage via the conjugate prior can be done for the finite GM as well, except that the relation (7) will be implicit instead of closed form.

Beyond its mathematical elegance and simplicity of use, we believe that an infinite model has practical importance as well. In many instances the number of items to rank is very large. Search engines come immediately to mind, understood as algorithms for retrieval by inexact matching from a large database. Under this umbrella fall not only the well-known web search engines, but also the various specialized algorithms for finding matches in biological data bases, like Sequest (Eng et al., 1994) and Blast (Altschul et al., 1990). These algorithms output ranked lists, from which the human user interprets only the top t entries. The data base, that is, the set of items, is usually not fixed; typically it is growing as more proteins, genes, web pages are discovered. It is natural under this scenario to assume that n is potentially infinite. As we have shown, this does not make working with the data more difficult, and occasionally makes it faster.

Acknowledgments

Thanks to Jon Wellner for bringing up the question of infinite permutations. This work was partially supported by NSF awards IIS-0313339 and IIS-0535100. A preliminary version of the IGM model was published as Meilă and Bao (2008).

Appendix A. Proofs

We give the proofs for the asymptotic results in the following subsection.

A.1 Proofs for the Asymptotic Results

Proof of Proposition 5 We start with the observation that under $P_{\mathrm{id},\theta}$ any observed top-t ranking π is a function of the variables $s_{1:t}$ who are independently distributed according to discrete exponential laws. For each s_j the empirical CDF converges to the true CDF and, as we know, this entails the fact that for any function f over $\mathbb{N}$, the sample expectation of f converges to the true expectation of f; see for example, van der Vaart (1998). This also holds for the joint distribution of $s_{1:t}$ and functions of $s_{1:t}$. In other words, the sample expectation of any function $f(\pi)$ is consistent, when π ranges over all top-t permutations.

Now consider $L_\sigma(R_j)$ for a fixed σ. By definition, $L_\sigma(R_j(\pi)) = s_j(\pi|\sigma)$ and $L_\sigma(\hat{R}_j)$ is the sample expectation of $s_j(\pi|\sigma)$. According to the definition of $s_j(\pi|\sigma)$ in (2), this is a function of $\pi^{-1}(1), \ldots \pi^{-1}(j)$ and the fixed σ. It follows by the argument above that $L_\sigma(\hat{R}_j)$ converges to $L_\sigma(R_j)$ for any j and any σ.

Note that for $\sigma = \text{id}$ the argument is simpler, since $L(R_j) = s_j$. □

We now refer to a property of the Mallows model introduced by Fligner and Verducci (1988). An IGM $P_{\sigma,\theta}$ has *complete consensus* if for any two items i, i' with $i \prec_\sigma i$, we have that $P[i \prec_\pi i'] > P[i' \prec_\pi i]$. The following result is a modified form of Theorem 2 of Fligner and Verducci (1988) that applies to truncated infinite permutations.

Proposition 11 (Complete consensus) *The IGM model $P_{\sigma,\theta}$ with*

$$\theta_j \;\geq\; \theta_{j+1}, \tag{22}$$

has complete consensus. Moreover, condition (22) entails that for any fixed t, any $j = 1 : t$, and any items i, i' with $\sigma(i) < \sigma(i')$

$$R_{ii',j} \;<\; R_{i'i,j}.$$

Proof Fix i, i' as above. Let π be a (complete) permutation where $i \prec_\pi i'$. Let us denote by π' the permutation obtained by transposing i and i' in π; denote $\pi(i) = k$, $\pi(i') = k'$, $k' > k$.

We want to show that under the condition of the proposition, $P_{\sigma,\theta}(\pi) \geq P_{\sigma,\theta}(\pi')$. We first observe that

$$s_j(\pi') = \begin{cases} s_j(\pi) & \text{for} \quad j < k \,\text{or}\, j > k' \\ s_j(\pi) + i' - i - r & \text{for} \quad j = k \,\text{and}\, r = |\{x \,|\, i < \sigma(x) < i', \pi(x) < k\}| \\ s_j(\pi) \,\text{or}\, s_j(\pi) + 1 & \text{for} \quad k < j < k' \\ s_j(\pi) - r' & \text{for} \quad j = k' \,\text{and}\, r' = |\{x \,|\, i < \sigma(x) < i', \pi(x) > k'\}| \end{cases} .$$

Note also that $r + r' \leq i' - i - 1$ or, in other words, $i' - i - r > r'$. Now, we look at the likelihood ratio $P_{\sigma,\theta}(\pi)/P_{\sigma,\theta}(\pi')$:

$$\begin{aligned} \ln[P_{\sigma,\theta}(\pi)/P_{\sigma,\theta}(\pi')] &= \sum_{j=k}^{k'} \theta_j [s_j(\pi') - s_j(\pi)], \\ &\geq \theta_k[s_k(\pi') - s_k(\pi)] + \theta_{k'}[s_{k'}(\pi') - s_{k'}(\pi)], \\ &= \theta_k(i' - i - r) - \theta_{k'} r', \\ &\geq \theta_k r' - \theta_{k'} r' \;=\; (\theta_k - \theta_{k'}) r' \;\geq\; 0. \end{aligned}$$

It follows that $P_{\sigma,\theta}(\pi) \geq P_{\sigma,\theta}(\pi')$. Moreover, if $\pi(i) = j, \pi(i') = j+1$ for some j then this inequality is strict. Now let $A = \{\pi \,|\, i \prec_\pi i'\}$ be the set of all permutations π as defined above; its complement B equals $\{\pi' \,|\, i' \prec_{\pi'} i\}$. It is immediate that from the above that $P_{\sigma,\theta}(A) = P_{\sigma,\theta}(i \prec i') > P_{\sigma,\theta}(B) = P_{\sigma,\theta}(i' \prec i)$, which proves the first claim of the proposition.

For the second claim, fix i, i' with $\sigma(i) < \sigma(i')$, a rank j and another rank $j' > j$. Take π such that $\pi(i) = j$ and $\pi(i') = j'$, and let π' be the permutation obtained by transposing i and i' in π. Then obviously $P_{\sigma,\theta}(\pi) > P_{\sigma,\theta}(\pi')$. $R_{ii',j}$, by definition, is the total probability of permutations of the form π with $j' = j+1, j+2, \ldots$, while $R_{i'i,j}$ is the total probability of permutations of the form π'. Therefore, $R_{ii',j} > R_{i'i,j}$. □

Proof of Proposition 5 From Proposition 11 it follows that if the true central permutation is the identity, then $L_{\mathrm{id}}(R_j) < L_\sigma(R_j)$ for any $\sigma \neq \mathrm{id}$. Let $\varepsilon = L_\sigma(R_j) - L_{\mathrm{id}}(R_j) > 0$. Then, because of the consistency of $L_\sigma(R_j)$ for any σ, it follows that $P_{\mathrm{id},\theta}[L_{\mathrm{id}}(\hat{R}_j) - L_\sigma(\hat{R}_j) > L_{\mathrm{id}}(R_j) - L_\sigma(R_j) + \varepsilon = 0] \to 0$. □

Proof of Proposition 6 This proof follows from the consistency of $L_\sigma(R_j)$. Since the ML estimate of θ_j is a continuous function of $L_\sigma(R_j)$ it will be consistent as well. Similarly, for the single parameter IGM, the ML estimate of θ is a continuous function of $(L_\sigma(R_j),\ j = 1:t)$, and therefore it is consistent as well. □

Proof of Proposition 10 Given the observed rankings $\{\pi_{1:N}\}$ and the hyperparameters $\nu, \Lambda_{1:t}$ the marginal distribution of the central permutation σ can be expressed in terms of S_j^*:

$$\begin{aligned} P(S_j^* = k) \propto Beta(k+1, N+1+\nu) &= \frac{\Gamma(k+1)\Gamma(N+1+\nu)}{\Gamma(k+N+2+\nu)} = f(k), \\ \frac{f(k+1)}{f(k)} = \frac{\Gamma(k+2)\Gamma(N+1+\nu)}{\Gamma(k+N+3+\nu)} \div \frac{\Gamma(k+1)\Gamma(N+1+\nu)}{\Gamma(k+N+2+\nu)} &= \frac{k+1}{N+k+2+\nu}, \\ f(k) = \frac{k!}{(N+2+\nu)\cdots(N+1+\nu+k)} \times f(0) &= \frac{k!}{(N+1+\nu)\cdots(N+1+\nu+k)}. \end{aligned}$$

We prove in Lemma 12 that $\sum_{k=0}^{\infty} f(k) = \frac{1}{(N+\nu)}$. Therefore, the normalization constant of $P(S_j^*)$ is $1/(N+\nu)$ and the conclusion of the Proposition follows. □

Lemma 12 $\sum_{k=0}^{\infty} \frac{k!}{(N+\nu+1)\cdots(N+\nu+1+k)} = \frac{1}{(N+\nu)}$.

Proof of Lemma 12 We write the general term of the series as a difference

$$\begin{aligned} \frac{1}{N+\nu} - \frac{1}{(N+\nu+1)} &= \frac{1!}{(N+\nu)(N+\nu+1)}, \\ \frac{1!}{(N+\nu)(N+\nu+1)} - \frac{1!}{(N+\nu+1)(N+\nu+2)} &= \frac{2!}{(N+\nu)(N+\nu+1)(N+\nu+2)}, \\ &\vdots \end{aligned}$$

Through mathematical induction we can prove that

$$\sum_{k=0}^{K} \frac{k!}{(N+\nu+1)\cdots(N+\nu+1+k)} + \frac{(K+1)!}{(N+\nu)(N+\nu+1)\cdots(N+K+\nu+1)} = \frac{1}{(N+\nu)}.$$

Moreover, for fixed N, ν,

$$\begin{aligned} \frac{(K+1)!}{(N+\nu)(N+\nu+1)\cdots(N+K+\nu+1)} &= \frac{(K+1)!(N+\nu-1)!}{(N+K+\nu+1)!}, \\ &= \frac{(N+\nu-1)!}{(K+2)(K+3)\cdots(N+K+\nu+1)}, \\ &= O(k^{-(N+\nu)}). \end{aligned}$$

From this the desired result follows. □

References

S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 3(213):403–410, 1990. doi:10.1006/jmbi.1990.9999., PMID 2231712.

L.M. Busse, P. Orbanz, and J. Bühmann. Cluster analysis of heterogeneous rank data. In *Proceedings of the International Conference on Machine Learning ICML*, 2007.

M.A. Carreira-Perpiñán. Fast nonparametric clustering with gaussian blurring mean-shift. In *23rd International Conference on Machine Learning (ICML)*, pages 153–160, 2006.

Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17 (8):790–799, 1995. ISSN 0162-8828. doi: http://dx.doi.org/10.1109/34.400568.

W.C. Cohen, R.S. Schapire, and Y. Singer. Learning to order things. *Journal of Artificial Intelligence Research*, 10:243–270, 1999.

D.E. Critchlow. *Metric methods for analyzing partially ranked data*. Number 34 in Lecture notes in statistics. Springer-Verlag, Berlin Heidelberg New York Tokyo, 1985.

M.H. DeGroot. *Probability and Statistics*. Addison–Wesley Pub. Co., Reading, MA, 1975.

J.K. Eng, A.L. McCormack, and J.R. Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society of Mass Spectrometry*, 5:976–989, 1994. doi:10.1016/1044-0305(94)80016-2.

M.A. Fligner and J.S. Verducci. Distance based ranking models. *Journal of the Royal Statistical Society B*, 48:359–369, 1986.

M.A. Fligner and J.S. Verducci. Multistage ranking models. *Journal of the American Statistical Association*, 88:892–901, 1988.

M.A. Fligner and J.S. Verducci. Posterior probability for a consensus ordering. *Psychometrika*, 55: 53–63, 1990.

K. Fukunaga and L.D. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Information Theory*, IT-21:32–40, 1975. ISSN 0018-9448.

K. Goldberg, T. Roeder, D. Gupta, and C. Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151, July 2001. http://goldberg.berkeley.edu/jester-data/, to cite for Jester data set.

I.C. Gormley and T.B. Murphy. Analysis of irish third-level college applications. *Journal of the Royal Statistical Society, Series A*, 169(2):361–380, 2006.

I.C. Gormley and T.B. Murphy. Exploring heterogeneity in irish voting data: A mixture modelling approach. Technical Report 05/09, Department of Statistics, Trinity College Dublin, 2005.

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

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

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

B. Mandhani and M. Meilă. Better search for learning exponential models of rankings. In David VanDick and Max Welling, editors, *Artificial Intelligence and Statistics AISTATS,* number 12, 2009.

Y. Mao and G. Lebanon. Non-parametric modelling of partially ranked data. *Journal of Machine Learning Research*, 9:2401–2429, 2008. URL `jmlr.csail.mit.edu/papers/v9/lebanon08a.html`.

M. Meilă and L. Bao. Estimation and clustering with infinite rankings. In David McAllester and Petri Millimäki, editors, *Proceedings of the 24-th Conference on Uncertainty in Artificial Intelligence (UAI 2008)*. AUAI Press, 2008.

M. Meilă, K. Phadnis, A. Patterson, and J. Bilmes. Consensus ranking under the exponential model. In Ron Parr and Linda Van den Gaag, editors, *Proceedings of the 23rd Conference on Uncertainty in AI*, volume 23, page (to appear), 2007.

T.B. Murphy and D. Martin. Mixtures of distance-based models for ranking data. *Computational Statistics and Data Analysis*, 41(3–4):645–655, 2003.

J. Pearl. *Heuristics: Intelligent Search Strategies for Computer Problem Solving*. Addison-Wesley, 1984.

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

N. Quadrianto, A.J. Smola, L. Song, and T. Tuytelaars. Kernelized sorting. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* (preprint), 2010.

R.P. Stanley. *Enumerative Combinatorics*, volume 1. Cambridge Unversity Press, Cambridge, New York, Melbourne, 1997.

E. Thoma. Die unzerlegbaren, positiv-definiten Klassenfunctionen der abzälbar unendlichen, symmetrische Gruppen. *Mathematische Zeitschrift*, 85:40–61, 1964.

A.W. van der Vaart. *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambrigde University Press, 1998.