Technical Report

Sampling and learning the Mallows and Weighted Mallows models under the Hamming distance

Ekhine Irurozki, Borja Calvo, José A. Lozano

January, 2014

Sampling and learning the Mallows and Weighted Mallows models under the Hamming distance

Ekhine Irurozki, Borja Calvo, Jose A. Lozano

Abstract

In this paper we deal with distributions over permutation spaces. The Mallows model is the mode l in use. The associated distance for permutations is the Hamming distance.

1   Introduction

Permutations appear in many areas and have been therefore studied for a long time. Moreover, probability distributions over permutations are also not new. The first models where introduced in the 20's while during the 50's they received the attention of the community. The most celebrated papers in the literature of this decade introduce the Placket-Luce [9], [15], Pairwise [**?**] and Mallows [10] models. Actually, these are the most lively models nowadays. One can find in the recent literature theoretical discussion and practical applications as well as extensions thereof. The main reason for this subject to be still an active topic is the interpretation of permutations as rankings and the big explosion on the number of applications of rankings. As examples of these applications we can mention preference elicitation, information retrieval, . . .

The probability model in this manuscript is the Mallows model (MM). Just two parameters are required to define such a model. The first one is a consensus permutation, σ_0, the one that receives the largest probability mass. The probability value of any other permutation $\sigma \in S_n$ decreases exponentially as the distance to the consensus permutation increases. The kurtosis of the distribution is controlled by the second parameter, θ, namely the spread parameter.

There are multiple extensions of the model. Some of the most popular extensions are non-parametric models [11], infinite permutations [8], [12] and mixture models [4], [13], [14]. The Generalized Mallows model (GMM) [7] is the most famous among all of them. Instead of one single spread parameter, it requires the definition of $n-1$ spread parameters θ_j for $1 \le j < n$, each affecting a particular position of the permutation. This allows modeling a distribution with more emphasis on the consensus of certain positions of the permutation while having more uncertainty in some others.

The original definition of the MM included two different metrics for permutations. However, in [5] this definition is extended to different metrics giving rise to the family of the distance based probability models. Six different distances were suggested, in particular Kendall's-τ, Hamming, Ulam, Cayley, Spearman's-ρ and Spearman's-footrule. However, the metric that has received more attention in the literature is the Kendall's-τ distance. The reason is that this metric is a natural way of measuring dissimilarities in the ranking or election domain. In this way, there are efficient algorithms for making inference, simulating and fitting a MM and GMM under the Kendall's-τ distance.

However, in many other domains the Kendall's-τ distance is not a natural way of measuring the differences between permutations. Instead, the most common metric in many areas is the Hamming distance (coding theory, cryptography, information theory, . . .). We believe that these powerful and simple models, MM and GMM, can be as useful in these areas as they are in the ranking domain. However, the basic operations for distributions (inference, sampling and learning) must be efficiently carried out. This paper tries thus to give a step forward in that direction by modeling the MM under the Hamming distance, in what we will refer to as Hamming Mallows (HM) and adapting the idea behind the GMM for the Hamming distance, the Weighted Hamming Mallows (WHM).

This paper is organized as follows. Section 3 introduces the probability models as well as the Hamming distance. Section 4 gives the closed formulas for the normalization constants of both MM and WHM. Section 5 exploits the TL-decomposability property in order to give an efficient method for the simulation of both MM and WHM. Section 6 deals with the estimation of the parameters of the distribution given a collection of permutations. In Section **??** we give efficient formulations for the algebraic machinery required in the rest of the paper, in particular, we include formulas for the computation of the elementary symmetric polynomials, their derivatives and for counting permutations by their number of unfixed points.

2   Introduction

Permutations are bijections of the set of integers $\{1, \ldots, n\}$ onto itself. They are usually denoted with the letters σ or π. From now on we will use the following notation, $\sigma(j) = i$ means that item i is in position j and represent the

permutation σ as $\sigma = [\sigma(1)\sigma(2)\ldots\sigma(n)]$. A special permutation which is worth mentioning is the identity permutation, $e = [123\ldots n]$ which maps each item j to position j.

By composing two permutations σ and π of n elements we obtain a new permutation $\sigma \circ \pi$ such that $\sigma \circ \pi(j) = \sigma(\pi(j))$, which will be denoted as $\sigma\pi$. The result of composing a permutation σ and its inverse σ^{-1} is the identity, $\sigma\sigma^{-1} = e$ and the composition $\sigma e = e\sigma = \sigma$.

The Hamming distance counts the number of disagreements between two strings. Its invariance property asserts that $d(\sigma, \pi) = d(\sigma\gamma, \pi\gamma)$ for every permutation γ. Particularly taking $\gamma = \pi^{-1}$ and since $\pi\pi^{-1} = e$ one can w.l.o.g. write $d(\sigma, \pi) = d(\sigma\pi^{-1}, e)$. The distance from any permutation to the identity is denoted as an univariate function $d(\sigma\pi^{-1}, e) = d(\sigma\pi^{-1})$ what will simplify the notation. The implications of this property are relevant since, as we will later explain, from now on we can w.l.o.g. assume that the reference permutation is the identity.

A recurrent concept in the permutation literature is that of fixed point. A fixed point is a position of the permutation in which $\sigma(i) = i$. In the same way, if $\sigma(i) \neq i$ then i an unfixed point. Therefore, the Hamming distance from a permutation σ to the the identity, $d(\sigma)$, counts the number of unfixed points of σ, $d(\sigma) = \sum_{j=1}^{n} I[\sigma(j) \neq j]$. Rather than the distance, we will sometimes be interested in the sets of fixed and unfixed points. This information is encoded on $h(\sigma)$, a binary vector defined as follows:

$$h_j(\sigma) = \begin{cases} 0 & \text{if } j \text{ is a fixed point} \\ 1 & \text{otherwise} \end{cases}$$

The $h(\sigma)$ vector is referred to as the distance decomposition since $d(\sigma) = \sum_{j=1}^{n} h_j(\sigma)$. We will use the denomination "fixed point at i", $\sigma(i) = i$ and $h_i(\sigma) = 0$ interchangeably along the paper. The notion of derangement is also relevant throughout this manuscript. A derangement is a permutation in which there are not fixed points.

2.1 Background

In this paper we deal with the efficient computation of operations for distributions over permutations. These operations, such as conditioning, sampling and learning, relay on operations for permutations, such as the random generation of permutations. Moreover, these basic operations for permutations are recurrent in the whole manuscript. Therefore, the efficient computation of operations for permutations is critical for the efficient management of distributions over permutations. For this reason, we first show how this key operations are carried out. These operations include counting permutations, random generation of permutations and the fast computation of the elementary symmetric polynomials of a set of variables.

2.1.1 Counting permutations at each possible distance

We deal first with the issue of counting the number of distinct permutations of n items at a given distance d. We will later use this sequence in the radom generation of permutations and in the sampling process. Note that in a permutation at Hamming distance d there are $n - d$ fixed points and d unfixed points. Therefore, we are interested on counting the number of permutations with exactly d unfixed points and exactly $n - d$ fixed points.

This question is closely related to the notion of derangement. As we have already stated, in a permutation at Hamming distance d there are d unfixed points, i.e., there are d items that form a derangement. Therefore, the number of permutations of n items at Hamming distance d is

$$S_h(n, d) = \binom{n}{d} S(d) \quad \text{where } S(d) \text{ is the number of derangements of } d \text{ items}$$

Counting the number of derangements is a recurring question and it is given in the well known On-Line Encyclopedia of Integer Sequences (OEIS) with code A000166. Unfortunately, there is no closed form for $S(d)$. The computation of $S(d)$ must be done in a recursive way as the following:

$$S(d) = \begin{cases} 1 & d = 0 \\ 0 & d = 1 \\ (d-1) * S(d-1) + (d-2) * S(d-1) & \text{otherwise} \end{cases}$$

This equation can be computed in $O(d)$. Since we are interested on the derangements of at most n items, we must compute $S(d)$ for $0 \leq d \leq n$ that requires time $O(n)$.

2.1.2 Counting permutations by unfixed points

We have already shown how to count the number of permutations with exactly d unfixed points and exactly $n - d$ fixed points in the previous lines. We deal now with the similar problem of counting the number of permutations with at least k unfixed points or with at least k fixed points. This count will be later useful for the closed form of the normalization constant of the probability model and in the learning process.

Note that the number of permutations such that have fixed points at positions $1 \le i_1 < i_2 < \ldots < i_k \le n$ is the same as the number of permutations that have fixed points at positions $1, 2, \ldots, k$. This count is denoted as $f(n, k)$. This fact simplifies the algebra since it is easy to see that $f(n, k) = (n - k)!$.

The same situation happens when we are counting the permutations with unfixed positions, i.e., the number of permutations such that have at least k unfixed points at positions $1 \le i_1 < i_2 < \ldots < i_k \le n$ is the same as the number of permutations that have unfixed points at positions $1, 2, \ldots, k$. This count is denoted as $g(n, k)$. We will compute this number using an inclusion-exclusion approach.

Counting $g(n, 1)$ The set of $n!$ permutations can be divided into two groups regarding $h_1(\sigma)$, those permutations in which $h_1(\sigma) = 0$ and those in which $h_1(\sigma) = 1$. The set of permutations in the former group is $f(n, 1) = (n - 1)!$. The cardinality of the latter set is therefore

$$g(n, 1) = n! - f(n, 1)$$

Counting $g(n, 2)$ The number of permutations in which $\sigma(1) \ne 1$ and $\sigma(2) \ne 2$ (i.e., those permutations in which both 1 and 2 are unfixed points) are computed in the same way, from the complete set of $n!$ permutations we discard those in which $\sigma(1) = 1$ ($f(n, 1)$) and then those in which $\sigma(2) = 2$ ($f(n, 1)$) -note that the number of permutations such that $\sigma(2) = 2$ is equal to the number of permutations such that $\sigma(1) = 1$, that is $f(n, 1)$.

However, if we set $g(n, 2) = n! - f(n, 1) - f(n, 1)$ we are under counting, since those permutations such that $\sigma(1) \ne 1 \wedge \sigma(2) \ne 2$ are substracted two times. Therefore, we must compensate this undercount by adding again those permutations that $\sigma(1) \ne 1 \wedge \sigma(2) \ne 2$

$$g(n, 2) = n! - f(n, 1) - f(n, 1) + f(n, 2)$$

Counting $g(n, l)$ The process continues in the same way. It is then easy to see that

$$g(n, k) = n! + \sum_{i=1}^{k} (-1)^i \binom{k}{i} f(n, i) = n! + \sum_{i=1}^{k} (-1)^i \frac{k!(n-i)!}{i!(k-i)!} \tag{1}$$

Therefore, with Equation (1) counting the number of permutations with a given number number of unfixed positions can be done in $O(n)$.

2.1.3   Random generation of permutations at a prescribed distance

In this section we deal with the random generation of permutations at a given distance. This technique will be used in the sampling algorithms. We will consider two different scenarios for the random generation of permutations.

- The set of fixed and unfixed points is given.

- Just the distance is given, but not the separate sets of fixed and unfixed points.

Note that a possible approach to generate a permutation of n items at Hamming distance d is to randomly select d unfixed points, make a derangement with them. The remaining $n - d$ items j can be placed so that $\sigma(j) = j$. In this way, both problems relay on randomly generating a derangement of d given items.

In order to generate random derangements we will use a recursive method again.

The base case is the generation of a derangement of two items, that will return the permutation $\sigma = [21]$.

For any other case, the generation of a derangement of d items implies first the generation of a derangement of $d - 1$ items and then the insertion of item d. The set of derangements of d items can be split in two groups: that which includes the derangements in which item d is in a cycle of length 2 and that which includes the derangements in which item d is in a cycle of length greater than 2 (recall that it can not be in a cycle of length one since a derangement implies that $\sigma(d) \ne d$). So to generate a derangement of d items we can either

- First, randomly select an item $0 < i < n$. Then, generate a derangement with items $\{1, \ldots, n - 1\}/\{i\}$. Finally set $\sigma(i) = d$ and $\sigma(d) = i$. In this way, item d in the resulting derangement is in a cycle of length 2. In this case, the recursion implies the generation of a derangements of the $S(n - 2)$ possible derangements of $d - 2$ items. Also, there are $d - 1$ possible ways of selecting item i. Therefore, there are exactly $(d - 2) * S(d - 1)$ derangements of d items of this form.

- First, generate a derangement with items $\{1, \ldots, n - 1\}$. Then, randomly select an item $0 < i < n$. Finally set $\sigma(i) = d$ and $\sigma(d) = i$. In this way, item d in the resulting derangement is in a cycle of length greater than 2. In this case, the recursion implies the generation of a derangements of the $S(n - 1)$ possible derangements of $d - 1$ items. Also, there are $d - 1$ possible ways of selecting item i. Therefore, there are exactly $(d - 1) * S(d - 1)$ derangements of d items of this form.

Therefore, of the total $S(d)$ derangements of d items, exactly $(d-2)*S(d-1)$ have item d in a cycle of length 2 and $(d-1)*S(d-1)$ of then have item d in a cycle of length greater than 2. So the probability of selecting the first branch is $(d-2)*S(d-1)/S(d)$ while the probability of the second is $1-(d-2)*S(d-1)/S(d) = (d-1)*S(d-1)/S(d)$.

The code of the whole generation process is given in Algorithm 1. Note that the complexity of generating a derangement of d items is $O(d)$.

Algorithm 1: $generate\_derangement(n,k)$

This algorithm generates a random derangement of n items. Note that every derangement of n items is equally probable.

Input: d, number of items;
Output: π, derangement of d items
if $d = 2$ **then** $\pi = [21]$; /* base case   */
else
    $prob = (d-2)*S(d-1)/S(d)$;
    with probability $prob$ /* d is in a cycle of length 2   */
        $\pi(\{1,\ldots,n-1\}/\{i\}) = generate\_derangement(d-1)$;
        $\pi(i) = n$;
        $\pi(n) = i$;
    end
    otherwise /* n is in a cycle of length grater than 2   */
        $\pi(1\ldots d-1) = generate\_derangement(d-1)$;
        ran = random number in the range $[1, n-1]$;
        $\pi(ran) = d$;
        $\pi(d) = ran$;
    end
end
return π;

2.1.4 Elementary symmetric polynomial

Every operation for the probability models of parameters $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_n)$ introduced in this paper relays on the computation of the Elementary Symmetric Polynomials (ESP) on the parameters of the distribution, $\boldsymbol{\theta}$. Therefore, the efficient computation of the ESP is crucial for the performance of the introduced algorithms. In this section we introduce the notion of ESP and show how to compute them efficiently. Moreover, we will give an expression for their derivatives.

The ESP of degree k in a set of n variables, $\gamma_k(X_1,\ldots,X_n)$, is defined as follows:

$$\gamma_k(X_1,\ldots,X_n) = \sum_{1 \le i_1 < \ldots < i_k \le n} \prod_{s=1}^{k} X_{i_s}$$

By abusing notation γ_k will be used to denote the ESP of degree k, $\gamma_k = \gamma_k(X_1,\ldots,X_n)$. The ESP γ_k can be efficiently computed with the following recursion, [1]:

$$\gamma_k(X_1,\ldots,X_n) = \begin{cases} 1 & \text{if k=0} \\ \sum_{i=1}^{n} X_i & \text{if k=1} \\ \gamma_k(X_1,\ldots,X_{n-1}) + \gamma_{k-1}(X_1,\ldots,X_{n-1})*X_n & \text{otherwise} \end{cases} \quad (2)$$

By using the above formula, the computational complexity for computing the ESP $\gamma_k(X_1,\ldots,X_n)$ is $O(n^2)$. Note that a naive computation will require time $O(2^n)$.

Splitting the ESP In the next lines there is an example of the elementary symmetric polynomial on 4 variables. Each γ_k is computed by adding up every products inside the braces.

$$\gamma_1 = \begin{cases} \gamma_1^1 = \{X_1 \\ \bar{\gamma}_1^1 = \begin{cases} X_2 \\ X_3 \\ X_4 \end{cases} \end{cases} \quad \gamma_2 = \begin{cases} \gamma_2^1 = \begin{cases} X_1X_2 \\ X_1X_3 \\ X_1X_4 \end{cases} \\ \bar{\gamma}_2^1 = \begin{cases} X_2X_3 \\ X_2X_4 \\ X_3X_4 \end{cases} \end{cases} \quad \gamma_3 = \begin{cases} \gamma_3^1 = \begin{cases} X_1X_2X_3 \\ X_1X_2X_4 \\ X_1X_3X_4 \end{cases} \\ \bar{\gamma}_3^1 = \{X_2X_3X_4 \end{cases} \quad \gamma_4 = \begin{cases} \gamma_4^1 = \{X_1X_2X_3X_4 \\ \bar{\gamma}_4^1 = \{0 \end{cases}$$

As we can see, the addends are divided into two groups. By γ_k^i we denote the subset of the addends in γ_k that include the term X_i, that is

$$\gamma_k^i = \sum_A \prod_{j \in A} X_j \text{ where } A \subseteq \{1, \ldots, n\} \wedge |A| = k \wedge i \in A$$

and by $\bar{\gamma}_k^i$ we denote the subset of the addends in γ_k that do not include the term X_i, that is

$$\bar{\gamma}_k^i = \sum_A \prod_{j \in A} X_j \text{ where } A \subseteq \{1, \ldots, n\} A \wedge |A| = k \wedge i \notin A$$

Some operations over the mentioned probability distributions require these two values to be given separately. Clearly, each of $\bar{\gamma}_k^i$ and γ_k^i for every $1 \leq i, k \leq n$ can be computed with Equation (2). However, we introduce a method for computing γ_k^i and $\bar{\gamma}_k^i$ for every $1 \leq i, k \leq n$ given γ_k in time $O(n^2)$. It is based in the following two relations:

$$\gamma_k = \gamma_k^i + \bar{\gamma}_k^i \text{ for every } i \tag{3}$$
$$\gamma_k^i = \bar{\gamma}_{k-1}^i X_i \text{ for every } i \tag{4}$$

By means of Equations (3) and (4) and being γ_k for $1 \leq k \leq n$ computed as shown in Equation (2), it is easy to define a recursive procedure to compute γ_k^i and $\bar{\gamma}_k^i$. Let $\gamma_0^i = \bar{\gamma}_0^i = 1$ and $\gamma_1^i = X_i$.

- The computation of $\bar{\gamma}_1^i$ is now trivial due to Equation (3), $\bar{\gamma}_1^i = \gamma_1 - \gamma_1^i$.

- Now, γ_2^i is easily computed by means of Equation (4), $\gamma_2^i = \bar{\gamma}_1^i X_i$.

- Then, Equation (3) can be used to compute $\bar{\gamma}_2^i$, $\bar{\gamma}_2^i = \gamma_2 - \gamma_2^i$.

In this way, the recursion follows until the γ_k^i and $\bar{\gamma}_k^i$ are computed for all $1 \leq i, k \leq n$ in $O(n^2)$.

Derivatives Since the variables in the elementary symmetric polynomial are exponential functions of the form $X_i = (exp(-\theta_i) - 1)$ the techniques for computing the derivatives explained in [1] are not valid here. We give here an efficient expression for the first and second derivatives. The first derivative of the ESP of degree k can be expressed as follows

$$\frac{\delta \gamma_k}{\delta \theta_i} = -exp(-\theta_i)\bar{\gamma}_{k-1}^i \tag{5}$$

The second derivative of the ESP of degree k can be expressed as follows

$$\frac{\delta^2 \gamma_k}{\delta \theta_i \theta_j} = \begin{cases} exp(-\theta_i)exp(-\theta_j)\bar{\gamma}_{k-2}^{i,j} & \text{if } i \neq j \\ exp(-\theta_i)\bar{\gamma}_{k-1}^i & \text{otherwise} \end{cases} \tag{6}$$

The proof, an example of the expansion of several ESP as well as the computation of a derivative are given in Appendix 7.

3   Mallows model, an extension thereof and hamming distance

The Mallows model (MM) is an exponential location probability model for permutations based on distances. It can be expressed as follows:

$$p(\sigma) = \frac{exp(-\theta d(\sigma, \sigma_0))}{\psi(\theta)}$$

where $\theta \in \mathbb{R}$ is a spread parameter, σ_0 is the location parameter called the central permutation, $d(\sigma, \sigma_0)$ represents a distance between σ and σ_0 and $\psi(\theta)$ the normalization constant $\psi(\theta) = \sum_\sigma exp(-\theta d(\sigma, \sigma_0))$. Note that when the dispersion parameter θ is greater than 0 then σ_0 is the permutation with the largest probability mass (the mode), and the closer a permutation σ is to σ_0, the larger $p(\sigma)$. On the other hand, with $\theta = 0$ we obtain the uniform distribution and when $\theta < 0$ then σ_0 is the anti mode. We will refer to the MM under the Hamming distance as Hamming Mallows (HM).

It is a common situation that the distances between permutations depend not only in the number of discrepancies but also in the positions of those discrepancies. Think for example in the ranking domain, where two permutations that disagree in the last two candidates of an election may be closer to each other than two rankings that disagree in both the first and the last candidate. Under the Kendall's-τ distance one can use the best know extension of the MM, the Generalized Mallows model (GMM) [7]. Instead of a single spread parameter it considers a vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{n-1})$ of $n-1$ spread parameters, each affecting a particular position of the permutation. The good news is that, under certain circumstances the GMM can be factored in $n-1$ terms [7]. Unfortunately, the GMM can not be coupled with

the Hamming distance. The reason is that in order to base the GMM on a particular distance, this distance must be decomposable as the sum on $n - 1$ terms, each related to a particular position of the permutation. Recall that the Hamming distance is decomposed in n terms.

However, the idea behind the GMM can be adapted to the use with Hamming distance by defining $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ spread parameters. We will refer to this model as Weighted Hamming Mallows (WHM) and define it as follows:

$$p(\sigma) = \frac{exp(-\sum_{j=1}^n \theta_j h_j(\sigma))}{\psi(\boldsymbol{\theta})}$$

where $\psi(\boldsymbol{\theta})$ the normalization constant $\psi(\boldsymbol{\theta}) = \sum_\sigma exp(-\sum_{j=1}^n \theta_j h_j(\sigma))$.

4   Inference

In order to calculate the probability value of a permutation under the MM or the WHM, the normalization constants must be computed. Given that the naive computation sums over $n!$ permutations this sum is an important bottle-neck when dealing with distributions over permutations of more than $n = 10$ items. Fortunately, there are closed forms for both normalization constants. The closed forms can be found by relating the normalization constants to the moment generating function of either the distance (in HM) or its decomposition vector (in WHM).

In particular, the process of finding the closed form of the normalization in the HM starts by expressing the probability generating function of the distance. In the WHM case, we will begin by defining the probability generating function of the decomposition vector of the distance. Then, we will find an alternative expression for the probability generating function by making use of a Taylor expansion. Finally, we will relate both probability generating function and moment generating function.

4.1   Hamming Mallows

The closed form of $\psi(\theta)$ under HM has already been given in [7]. We follow the same procedure used by them but using a slightly different notation: we use the reverse notation for $h_j(\sigma)$ since this will be helpful for the computation of the normalization constant in WHM. Recall that for us $h_j(\sigma) = 0$ iff j is a fixed point. Let σ be a u.a.r. chosen permutation and its Hamming distance $X = d(\sigma)$ a random variable. Let $P_0(d(\sigma) = d)$ be the probability of an u.a.r. chosen permutation σ of having $d(\sigma) = d$. The normalization constant of the HM can be posed as a function of the moment generating function of the distances as follows:

$$\psi(\theta) = n! \sum_d P_0(d(\sigma) = d)exp(-d\theta) = n!M(-\theta)$$

where the moment generating function $M(-\theta) = \sum_d P_0(d(\sigma) = d)exp(-d\theta)$.

The probability generating function of X is $f_x(t) = E[t^X]$. Its k-th derivative evaluated at $t = 1$ is

$$f_x^{(k)}(1) = E[X(X-1)(X-2)\ldots(X-(k-1))] = E\left[\frac{X!}{(X-k)!}\right]$$

So the expansion of $f(t)$ in a Taylor series around $t = 1$ is

$$f_x(t) \sum_{k=0}^\infty \frac{f^{(k)}(1)}{k!}(t-1)^k = \sum_{k=0}^n E[C(X,k)](t-1)^k$$

where the binomial coefficient $C(X, k)$ is 0 when $X < k$. Let $A = \{(i_1, \ldots, i_k)|1 \le i_1 < i_2 < \ldots < i_k \le n\}$.

For every $\mathbf{h} = (h_1, \ldots, h_n) \in \{0, 1\}^n$ the binomial coefficient may be re-expressed as

$$C(X, k) = \sum_A h_{i_1} \cdots h_{i_k}$$

Consequently, if we consider that $h_j = h_j(\sigma) = I[\sigma(j) \ne j]$

$$E[C(X,k)] = \sum_A E[h_{i_1} \cdots h_{i_k}] = \sum_A \frac{g(n,k)}{n!} = \frac{n!g(n,k)}{(n-k)!n!k!} = \frac{g(n,k)}{(n-k)!k!}$$

where $g(n, k)$ denotes the number of permutations such that have unfixed points at positions $1 \le i_1 < i_2 < \ldots < i_k \le n$ (which is the same as the number of permutations that have unfixed points at positions $1, 2, \ldots, k$, see Section (2.1.2) for details).

Going back to the Taylor expansion, $f(t)$ can be re-formulated as:

$$f(t) = \sum_{k=0}^{n} \frac{g(n,k)}{(n-k)!k!}(t-1)^k$$

Since $f(t) = E[t^X]$ then $f(exp(t)) = E[exp(t)^X]$. Now, the moment generating function $M(t) = E[exp(Xt)] = E[exp(t)^X] = f(exp(t))$. Therefore,

$$M(t) = f(exp(t)) = \sum_{k=0}^{n} \frac{g(n,k)}{(n-k)!k!}(exp(t)-1)^k$$

and therefore

$$M(-\theta) = \sum_{k=0}^{n} \frac{g(n,k)}{(n-k)!k!}(exp(-\theta)-1)^k$$

Finally, the normalization constant can be written as follows:

$$\psi(\theta) = n!M(-\theta) = n!\sum_{k=0}^{n} \frac{g(n,k)}{(n-k)!k!}(exp(-\theta)-1)^k \qquad (7)$$

Given $g(n,k)$ and the factorial numbers $k!$, which are both computed in $O(n)$, the computational complexity of Equation (7) is $O(n)$.

4.2 Weighted Hamming Generalized Mallows

By using the same reasoning and notation than in the HM case, we will now focus on the closed form expression of the normalization constant $\psi(\boldsymbol{\theta})$ for the WHM. Let $\mathbf{h} = (h_1, \ldots, h_n)$ and recall that $h_j \in \{0, 1\}$ and $h_j(\sigma) = 0$ iff $\sigma(j) = j$. Then $P_0(h(\sigma) = \mathbf{h})$ is the probability of an u.a.r. chosen permutation σ of having the distance decomposition $\mathbf{h}$. Being the multivariate (joint) moment generating function of the random variable $\mathbf{X}$ defined as $M_X(\mathbf{t}) = E[\prod_{j=1}^{n} exp(t_j X_j)]$, the normalization constant, $\psi(\boldsymbol{\theta})$, can be posed as follows

$$\psi(\boldsymbol{\theta}) = n! \sum_{\mathbf{h} \in \{0,1\}^n} P_0(h(\sigma) = \mathbf{h})exp(-\sum_j \theta_j h_j) = n! \sum_{\mathbf{h} \in \{0,1\}^n} P_0(h(\sigma) = \mathbf{h}) \prod_{j=1}^{n} exp(-\theta_j h_j) = n!M(-\boldsymbol{\theta})$$

The multivariate case of the probability generating function is defined as:

$$f(\mathbf{t}) = f(t_1, \ldots, t_n) = E[t_1^{h_1} \cdots t_n^{h_n}] = \sum_{\mathbf{h} \in \{0,1\}^n} P_0(h(\sigma) = (h_1, \ldots, h_n))t_1^{h_1} \cdots t_n^{h_n}$$

The Taylor expansion of a multivariate function is

$$f(\mathbf{t}) = \sum_{k=0}^{\infty} \frac{1}{k!} \sum_{x_1+\cdots+x_n=k} \binom{k}{x_1 \cdots x_n} \frac{\delta^k f}{\delta t_1^{x_1} \ldots \delta t_n^{x_n}}\bigg|_{\mathbf{t}=\mathbf{1}} (t_1-a_1)^{x_1} \cdots (t_n-a_n)^{x_n}$$

The derivative for variable t_i:

$$\frac{\delta f}{\delta t_i} = \sum_{(h_1,\ldots,h_n)} P_0(h(\sigma) = (h_1, \ldots, h_n))t_1^{h_1} \cdots h_i t_i^{h_i-1} \cdots t_n^{h_n}$$

$$= \sum_{(h_1,\ldots,h_n)} P_0(h(\sigma) = (h_1, \ldots, h_n) \wedge h_i = 0)t_1^{h_1} \cdots 0 t_i^{0-1} \cdots t_n^{h_n} + \sum_{(h_1,\ldots,h_n)} P_0(h(\sigma) = (h_1, \ldots, h_n) \wedge h_i = 1)t_1^{h_1} \cdots 1 t_i^{1-1} \cdots t_n^{h_n}$$

$$= 0 + \sum_{(h_1,\ldots,h_n)} P_0(h(\sigma) = (h_1, \ldots, h_n) \wedge h_i = 1) \prod_{j \neq i} t_j^{h_j}$$

that is, make two different groups, one for the permutations such that $h_i = 0$ and the other for the permutations $h_i = 1$. And its evaluation around $\mathbf{t} = (1, \ldots, 1)$ is :

$$\frac{\delta f}{\delta t_i}\bigg|_{\mathbf{t}=\mathbf{1}} = \sum_{(h_1,\ldots,h_n)} P_0(h(\sigma) = (h_1, \ldots, h_n) \wedge h_i = 1)1^{h_1} \cdots 1 \cdot 1^{1-1} \cdots 1^{h_n}$$

Note that this is equivalent to the probability under the uniform distribution of a permutation σ s.t. $\sigma(i) \neq i$ and so $h_i = 1$, i.e. the number of permutations of n with an unfixed point in i divided by $n!$. The second order derivative with

7

respect to t_i equals 0. The second order cross partial derivatives equal the probability under the uniform distribution of a permutation σ in which i_1 and i_2 are unfixed points, i.e. the number of permutations of n with unfixed points in i_1 and i_2 divided by $n!$

$$\frac{\delta^2 f}{\delta t_{i_1} t_{i_2}}\Big|_{\mathbf{t}=\mathbf{1}} = \sum_{(h_1,\ldots,h_n)} P_0(h(\sigma) = (h_1,\ldots,h_n) \wedge h_{i_1} = 1, h_{i_2} = 1)1^{h_1}\cdots 1 \cdot 1^{1-1}\cdots 1^{h_n}$$

Let $g(n,k)$ denote the number of permutations of n items with at lest k unfixed points. Recall than an efficient expression for the computation of $g(n,k)$ is given in Section (2.1.2). Then, in general, the k-th order cross partial derivatives equal:

$$\frac{\delta^k f}{\delta t_{i_1}\ldots t_{i_k}}\Big|_{\mathbf{t}=\mathbf{1}} = \frac{g(n,k)}{n!}$$

Since $\mathbf{h} = (h_1,\ldots,h_n) \in \{0,1\}^n$ then $\binom{k}{h_1\cdots h_n} = k!$. Since $\delta^k f/\delta t_i^k = 0$ for $k > 1$ the Taylor series cannot be expanded more than $n+1$ terms. Recall that $A = \{(i_1,\ldots,i_k)|1 \le i_1 < i_2 < \ldots < i_k \le n\}$. Therefore, the Taylor expansion around $\mathbf{a} = \mathbf{1} = (1,\ldots,1)$ can be equivalently written as:

$$f(\mathbf{t}) = \sum_{k=0}^{n} \frac{1}{k!} \sum_A \binom{k}{h_1\cdots h_n} \frac{\delta^k f}{\delta t_{i_1}\ldots t_{i_k}}\Big|_{\mathbf{t}=\mathbf{1}} \prod_{s=1}^{k}(t_{i_s} - 1) = \sum_{k=0}^{n} \frac{1}{k!} \sum_A \frac{k!g(n,k)}{n!} \prod_{s=1}^{k}(t_{i_s} - 1)$$

$$= \sum_{k=0}^{n} \frac{k!g(n,k)}{n!k!} \sum_A \prod_{s=1}^{k}(t_{i_s} - 1)$$

By setting $T_j = t_j - 1$

$$f(\mathbf{t}) = \sum_{k=0}^{n} \frac{g(n,k)}{n!} \sum_A \prod_{s=1}^{k} T_{i_s}$$

If γ_k denotes the Elementary Symmetric Polynomial, $\gamma_k(X_1,\ldots,X_n) = \sum_{1 \le i_1 < \ldots < i_k \le n} X_{i_1}\ldots X_{i_k}$:

$$f(\mathbf{t}) = \sum_{k=0}^{n} \frac{g(n,k)}{n!} \gamma_k(T_1,\ldots,T_n)$$

An efficient formulation for the computation of Elementary Symmetric Polynomial, $\gamma_k(X_1,\ldots,X_n)$ is given in Section 2.1.4. Note that $M(-\boldsymbol{\theta}) = E[\prod_{j=1}^{n} exp(-\theta_j h_j)] = E[\prod_{j=1}^{n} exp(-\theta_j)^{h_j}]$ and $f(\mathbf{t}) = E[\prod_{j=1}^{n} t_j^{h_j}]$, so $f(exp(-\boldsymbol{\theta})) = M(-\boldsymbol{\theta})$, and therefore

$$M(-\boldsymbol{\theta}) = \sum_{k=0}^{n} \frac{g(n,k)}{n!} \gamma_k(T_1,\ldots,T_n)$$

where $T_j = exp(-\theta_j) - 1$.

The normalization constant can thus be given as follows:

$$\psi(\boldsymbol{\theta}) = n!M(-\boldsymbol{\theta}) = \sum_{k=0}^{n} g(n,k)\gamma_k(T_1,\ldots,T_n) \tag{8}$$

where $T_j = exp(-\theta_j) - 1$.

Given $\gamma_k(T_1,\ldots,T_n) \forall k$, which are computed in $O(n^2)$, and $g(n,k)$ which requires $O(n)$ the computational complexity of Equation (8) is $O(n)$.

Properties of the models   In [2] some frequent properties of probability models for permutation domains are defined. Regarding the HM and WHM models, we can state that both are label invariant. They are not reversible, neither strongly unimodal and do not have complete consensus. On the other hand, both are L-decomposable as well as TL-decomposable. In fact, the TL-decomposable models are a subset of the L-decomposable models.

On [3] it is stated that a probabilistic model is TL-decomposable iff it is quasi-independent, i.e. it can be written as follows:

$$p(\sigma) = k \prod_{j=1}^{n} c_{j,\sigma(j)}$$

A quasi independent probability model as the one above requires n^2 parameters. We state that the WHM is also a TL-decomposable model. Moreover, the WHM belongs to a restriction of this class of models because it makes use

of just $2n$ parameters. Let us show this point in detail. Under the WHM the probability of a particular permutation σ is

$$p(\sigma) = exp(\sum_{j=1}^{n} -\theta_j h_j(\sigma \sigma_0^{-1}))\psi^{-1} = \psi^{-1}(\boldsymbol{\theta}) \prod_{j=1}^{n} exp(-\theta_j h_j(\sigma \sigma_0^{-1}))$$

Therefore, we can take $k = \psi^{-1}(\boldsymbol{\theta})$ and

$$c_{j,\sigma(j)} = \begin{cases} exp(0) = 1 & \text{if } h_j(\sigma \sigma_0^{-1}) = 0 \\ exp(-\theta_j) & \text{if } h_j(\sigma \sigma_0^{-1}) = 1 \end{cases}$$

We can conclude that both WHM and HM are TL-decomposable and consequently, L-decomposable. On an L-decomposable probability model the process of generating a permutation can be done as follows: The set of items to rank is given, each with an associated weight. At the first stage the most preferred item is randomly chosen, where the choice probability of each item is proportional to its weight. At the second stage, the most preferred item is chosen among the remaining items, where again, the choice probability of selecting each item is proportional to its weight. The generation process continues in this way, by setting the choice probability of the items at each stage regarding the set of remaining items and irrespective of the items that have been already chosen or the order in which they were selected. The process finishes when a complete permutation is built.

The L-decomposability property implies that the probability of a permutation is the product of the choice probabilities across the different stages, [2]. This means that the process of generating a permutation can be efficiently done if the choice probabilities are known.

4.3 Marginal probabilities

In this section we deal with the computation of the marginal probabilities. The marginal distribution is defined as the probability of a subset of variables irrespective of the values of the other variables. For the case of discrete random variables the marginal distribution is usually computed by summing the joint probability distribution over the unknown variables.

We are interested in the marginal distribution under the HM and WHM, in the permutation domain. Since the HM is a particular case of the WHM in which every θ_j has equal value, the rest of the section considers just the WHM model. Nevertheless, the results can be applied for both HM and WHM models. In particular, we are interested in questions such as "Which is the probability of a random permutation of having fixed point at i and an unfixed point at j?". As in other domains, this probability can be computed by summing the probabilities of every permutation having $\sigma(i) = i$ and $\sigma(j) \neq j$. Unfortunately, due to the factorial nature of the permutation space, this approach is infeasible.

In this section, we introduce a method for computing the marginal distribution. In order for this computation to be efficient, we avoid summing over every permutation. Its quick computation is based on the closed form expression for the normalization constant given in Equation (8) (Section 4.2).

Throughout this section we consider two sets of permutations. The first one is $fix(i)$, which includes every permutation in which i is a fixed point. By abusing notation we will also denote $fix(\{i_1, i_2\})$ as the set of permutations with fixed points at positions i_1, i_2. We define the set $unfix(i)$ in the same way, as the set of permutations that have an unfixed point at position i. Moreover, we will consider two sets of items A and B. In the rest of this section we show how to sum the probabilities of those permutations in which j is a fixed point for every $j \in A$ and j is an unfixed point for every $j \in B$. In particular, we are interested on the set of permutations in the intersection of the sets $fix(A)$ and $unfix(B)$, denoted as $fix(A) \cap unfix(B)$, and denote this marginal as follows:

$$\sum_{\sigma \in fix(A) \cap unfix(B)} p(\sigma) \tag{9}$$

As we have already stated, Equation (8) (Section 4.2) shows that the normalization constant can be efficiently computed as follows:

$$\psi(\boldsymbol{\theta}) = \sum_{\sigma} exp(\sum_{j=1}^{n} -\theta_j h_j(\sigma)) = \sum_{k=0}^{n} g(n,k)\gamma_k(T_1, \ldots, T_n)$$

Note that this equation sums $exp(\sum_{j=1}^{n} -\theta_j h_j(\sigma))$ over every permutation σ of n items. We will show how to adapt this idea to sum $exp(\sum_{j=1}^{n} -\theta_j h_j(\sigma))$ over every permutation σ' in the subset of permutations of interest. We begin by summing $exp(\sum_{j=1}^{n} -\theta_j h_j(\sigma))$ for every permutation of n items that has a fixed point at i. Note that if σ' ranges in the permutations $\sigma' \in fix(i)$ then it ranges over every possible permutation of the items $\{1, \ldots, n\} \setminus \{i\}$.

We consider two WHM distributions. The first one is over permutations of n items and has dispersion parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$. The second one is over distributions of $n-1$ items and has dispersion parameters $\boldsymbol{\theta}' = (\theta'_1, \ldots, \theta'_{n-1})$

which is equal to $\boldsymbol{\theta}$ in all but one parameter, θ_i, which has been removed. Also, let $T_j = exp(-\theta_j) - 1$ and $T'_j = exp(-\theta'_j) - 1$.

$$\sum_{\sigma \in fix(i)} exp(\sum_{j=1}^{n} -\theta_j h_j(\sigma)) = \sum_{\sigma \in fix(i)} exp(-\theta_i 0 + \sum_{j \neq i}^{n} -\theta_j h_j(\sigma)) = \sum_{\sigma'} exp(\sum_{j=1}^{n-1} -\theta'_j h_j(\sigma'))$$

$$= \sum_{k=0}^{n-1} g(n-1,k)\gamma_k(T'_1, \ldots, T'_{n-1}) = \sum_{k=0}^{n-1} g(n-1,k)\bar{\gamma}_k^i(T_1, \ldots, T_n) \qquad (10)$$

The first equality is based on the fact that if i is a fixed point then $h_i(\sigma) = 0$. The second one takes into account that for every $\sigma \in fix(i)$ a correspondence to permutations in $\sigma' \in S_{n-1}$ can be defined such that:

$$\mathbf{h}(\sigma) \setminus \{h_i\} = (h_1(\sigma), \ldots, h_{i-1}(\sigma), h_{i+1}(\sigma), \ldots, h_n(\sigma)) = \mathbf{h}(\sigma') = (h_1(\sigma'), \ldots, h_{n-1}(\sigma'))$$

From the third equality, we just applied Equation (8). Since $\boldsymbol{\theta} = \boldsymbol{\theta}' \cup \{\theta_i\}$, then $\bar{\gamma}_k^i(T_1, \ldots, T_n) = \gamma_k(T'_1, \ldots, T'_{n-1})$ the last equality holds.

The marginal distribution of the set of permutations with a fixed point at i is now trivial to compute.

$$\sum_{\sigma \in fix(i)} p(\sigma) = \frac{\sum_{\sigma \in fix(i)} exp(\sum_{j=1}^{n} -\theta_j h_j(\sigma))}{\psi(\boldsymbol{\theta})} = \frac{\sum_{k=0}^{n-1} g(n-1,k)\bar{\gamma}_k^i(T_1, \ldots, T_n)}{\sum_{k=0}^{n} g(n,k)\gamma_k(T_1, \ldots, T_n)}$$

Note that the naive computation of the marginal implies summing over $(n-1)!$ values while this alternative formulation has complexity $O(n^2)$, the same as the computing $g(n,k)$, γ_k and $\bar{\gamma}_k^i$. Recall that $\bar{\gamma}_k^i$ was introduce in Section 2.1.4. Moreover, in the same section, it is shown how to obtain $\bar{\gamma}_k^i$ given γ_k in $O(n^2)$ for every i, k.

The the second order marginal is computed with the same reasoning:

$$\sum_{\sigma \in fix(i_1,i_2)} p(\sigma) = \frac{\sum_{\sigma \in fix(i_1,i_2)} exp(\sum_{j \neq i_1,i_2}^{n} -\theta_j h_j(\sigma))}{\psi(\boldsymbol{\theta})} = \frac{\sum_{k=0}^{n-2} g(n-2,k)\bar{\gamma}_k^{i_1,i_2}(T_1, \ldots, T_n)}{\psi(\boldsymbol{\theta})}$$

In general, the probability of every $j \in A$ being a fixed point can be computed as follows.

$$\sum_{\sigma \in fix(A)} p(\sigma) = \frac{\sum_{\sigma \in fix(A)} exp(\sum_{j \notin A}^{n} -\theta_j h_j(\sigma))}{\psi(\boldsymbol{\theta})} = \frac{\sum_{k=0}^{n-|A|} g(n-|A|,k)\bar{\gamma}_k^A(T_1, \ldots, T_n)}{\psi(\boldsymbol{\theta})}$$

So far, we have considered the computation of the marginal probabilities of permutations with fixed points at positions j for all $j \in A$. We will now show how to compute the marginal probabilities of a permutations with fixed and unfixed points. In general, we are give two sets of integers A and B and we want to compute the marginal of those permutations having fixed points at positions j for all $j \in A$ and unfixed points at positions j for all $j \in B$. Since there is not a closed form for this expression, we propose an inclusion-exclusion approach to compute it. The idea is to compute first the probability of the set with fix points at A. Note that this group includes permutations that have fixed points at B as well as permutations that have unfixed points at B. The next step is therefore to remove the marginal probabilities of the permutations with fixed points at B.

We will begin the explanation of the computation of the marginal with just one unfixed point. Let B be another set of indexes such that $|B| = 1$. It is easy to see that

$$fix(A) \cap unfix(B) = fix(A) \setminus fix(A \cup B)$$

Therefore, the marginal probability of the permutations that have fixed points at positions A and unfixed points at position $i \in B$ can be given as follows:

$$\sum_{\substack{\sigma \in fix(A) \\ \cap unfix(B)}} p(\sigma) = \sum_{\sigma \in fix(A)} p(\sigma) - \sum_{\sigma \in fix(A \cup B)} p(\sigma)$$

However, when the set B gets larger, things get challenging. Let $B = \{i_1, i_2\}$ have two items. If we set the marginal of $fix(A) \cap unfix(B)$ as in previous example, $\sum_{\sigma \in fix(A)} p(\sigma) - \sum_{\sigma \in fix(A \cup B)} p(\sigma)$, then we are subtracting the probability of those permutations that have fixed points on A and i_1 and i_2 simultaneously. However, the probability of those permutations in which i_1 is a fixed point and i_2 is an unfixed point (and vice versa) are not subtracted.

On the other hand, in case we compute the marginal as $\sum_{\sigma \in fix(A)} p(\sigma) - \sum_{\sigma \in fix(A \cup \{i_1\})} p(\sigma) - \sum_{\sigma \in fix(A \cup \{i_2\})} p(\sigma)$, then the probability of the permutations $\sigma \in fix(i_1, i_2)$ are subtracted twice, one per sum. We are thus under counting. This can be solved by adding the probabilities of the permutations $\sigma \in fix(i_1, i_2)$ as follows:

$$\sum_{\substack{\sigma \in fix(A) \\ \cap unfix(B)}} p(\sigma) = \sum_{\sigma \in fix(A)} p(\sigma) - \sum_{\substack{\sigma \in \\ fix(A \cup \{i_1\})}} p(\sigma) - \sum_{\substack{\sigma \in \\ fix(A \cup \{i_2\})}} p(\sigma) + \sum_{\sigma \in fix(A \cup B)} p(\sigma)$$

In general, given sets A and B, the marginal distribution of those permutations having fixed points at positions j for all $j \in A$ and unfixed points at positions j for all $j \in B$ is as follows:

$$\sum_{\substack{\sigma \in fix(A) \\ \cap unfix(B)}} p(\sigma) = \sum_{i=0}^{|B|} (-1)^i \sum_{\substack{B' \subseteq B \\ |B'|=i}} \sum_{\substack{\sigma \in \\ fix(A \cup B')}} p(\sigma) = \sum_{i=0}^{|B|} (-1)^i \sum_{\substack{B' \subseteq B \\ |B'|=i}} \frac{\sum_{k=0}^{n-|A \cup B'|} g(n - |A \cup B'|, k) \bar{\gamma}_k^{A \cup B'}(T_1, \ldots, T_n)}{\psi(\boldsymbol{\theta})} \quad (11)$$

The computational complexity of this expression is $O(2^{|}B|)$.

4.4 Conditional probabilities

In this section we show how to compute conditional distribution under the HM and WHM models. Since the HM is a particular case of the WHM in which every dispersion parameter has equal value, we focus en the WHM. Nevertheless, we remark that these results can be applied to both models.

The conditional distribution gives the probability of a variable conditioned to a particular value of a (set of) variables. The conditional distribution can posed in terms of the joint distribution as follows:

$$p(X|Y) = \frac{X \cap Y}{p(Y)}$$

For our WHM distribution on permutation domains, we are interested on questions such as "which is the probability of item i being a fixed point given that item j is not a fixed point?". We will denote the above question as $p(h_i(\sigma) = 0|\sigma \in unfix(j))$. In order to pose the conditional distribution in terms of joint distribution we must compute the joint probability of the permutations such $h_i(\sigma) = 0$ and $\sigma \in unfix(j)$. Note that the set of permutations having both a fixed point in i and an unfixed point in j is $fix(i) \cap unfix(j)$, and its marginal distribution can be computed by means of Equation (11). Therefore, the conditional distribution can posed in terms of the joint distribution as follows:

$$p(h_i(\sigma) = 0|\sigma \in unfix(j)) = \frac{\sum_{\sigma \in fix(i) \cap unfix(j)} p(\sigma)}{\sum_{\sigma \in unfix(j)} p(\sigma)}$$

In general, we are given a set of fixed points A and a set of unfixed points, B. The probability of item i being a fixed point given that the items in A are fixed points and the items in B are unfixed points is given by the following expression:

$$p(h_i(\sigma) = 0|\sigma \in fix(A) \cap unfix(B)) = \frac{\sum_{\sigma \in fix(A \cup \{i\}) \cap unfix(B)} p(\sigma)}{\sum_{\sigma \in fix(A) \cap unfix(B)} p(\sigma)}$$

The computational complexity is thus equivalent to the complexity of the marginal computation.

5 Sampling

In this section we show how to generate permutations from both HM and WHM models. We introduce here three sampling algorithms. Two of them can generate from both WHM and HM in both exact and approximate manner, while the third algorithm can only generate permutations under the HM model in an exact way.

5.1 Marginal sampling

We propose a method for generating permutations from a HM or a WHM based on the marginal probabilities. It can be divided in two stages. In the first stage a binary vector is randomly generated. This vector corresponds to the distance decomposition of a permutation, $h(\sigma)$. The sampling finishes by randomly generating a permutation σ consistent with the given $h(\sigma)$.

The generation of the random distance decomposition vector, $h(\sigma)$, relays on the computation of the marginal and conditional probabilities, which are given in Sections 4.3 and 4.4 respectively. The generation of $h(\sigma)$ can be done in n stages. In the first stage $h_1(\sigma)$ is randomly selected using the marginal probabilities $\sum_{\sigma|\sigma(1)=1} p(\sigma)$ and $\sum_{\sigma|\sigma(1)\neq1} p(\sigma)$. Then, the probability of $h_2(\sigma)$ conditioned to the previous choice of h_1 is computed and the value of $h_2(\sigma)$ is randomly chosen. This process continues in the same way until the n positions of the vector are given. Before explicitly giving the explicit formula in the general case, we will show how the probabilities are computed for the first positions of the permutation.

The probability of the first position being a fixed point is easily computed given the marginal probabilities:

$$p(h_1(\sigma) = 0) = \sum_{\sigma | \sigma(1) = 1} p(\sigma) = \sum_{\sigma | \sigma \in fix(\{1\})} p(\sigma)$$

$$p(h_1(\sigma) = 1) = 1 - p(h_1(\sigma) = 0)$$

One of the options is randomly chosen. Throughout the process, sets A and B keep record of the sets of fixed and unfixed points respectively. Therefore, if $h_1(\sigma) = 0$ then $A = \{h_1\}, B = \emptyset$. Otherwise, $A = \emptyset, B = \{h_1\}$. In this way, we can compute the probability of the second position being a fixed point conditioned on the first position being (or not) a fixed point.

$$p(h_2(\sigma) = 0 | \sigma \in fix(A) \cap unfix(B)) = \frac{\sum_{\sigma \in fix(A \cup \{2\}) \cap unfix(B)} p(\sigma)}{\sum_{\sigma \in fix(A) \cap unfix(B)} p(\sigma)}$$

$$p(h_2(\sigma) = 1 | \sigma \in fix(A) \cap unfix(B)) = 1 - p(h_2(\sigma) = 0 | \sigma \in fix(A) \cap unfix(B))$$

Note that the value in the denominator is the probability selected in the previous stage. In general, the generation of the $h(\sigma)$ distance decomposition vector from a given HM or WHM distribution is done in n iterations. At each iteration i, $h_i(\sigma)$ is sampled. Let the sets A and B initially empty, each iteration i proceeds as follows:

- Compute the probabilities of being i a fixed point as follows:

$$p(h_j(\sigma) = 0 | \sigma \in fix(A) \cap unfix(B)) = \frac{\sum_{\sigma \in fix(A \cup \{j\}) \cap unfix(B)} p(\sigma)}{\sum_{\sigma \in fix(A) \cap unfix(B)} p(\sigma)}$$

$$p(h_j(\sigma) = 1 | \sigma \in fix(A) \cap unfix(B)) = 1 - p(h_j(\sigma) = 0 | \sigma \in fix(A) \cap unfix(B))$$

- Randomly choose $h_i(\sigma)$ given the previous probabilities.

- Update set A and B with the set of fixed and unfixed points.

In this way, our proposed sampling algorithm generates a binary vector $h(\sigma)$ in n stages. The process finishes by generating a permutation σ consistent with $h(\sigma)$. We refer the reader to Section 2.1.3 where the random generation of a permutation given the subsets of items that are fixed or unfixed is given. As a summary, we can state that this algorithm can generate permutations under both HM and WHM in an exact way. The complexity is equal to the complexity of the marginalization process.

5.2 Gibbs sampling

The Gibbs sampler is a Markov Chain Montecarlo algorithm based on sampling a Markov chain whose stationary distribution is the distribution of interest. Therefore, it is an approximated algorithm for the simulation of the distribution. We have adapted this algorithm to generate samples for both HM and WHM.

The Gibbs algorithm proceeds as follows:

1. Generate uniformly at random a permutation σ.

2. Build a new permutation σ' equal to σ in all but two positions, that have been swapped.

3. Let $\gamma = min\{1, p(\sigma')/p(\sigma)\}$. With probability γ the algorithm accepts the candidate permutation moving the chain to the candidate permutation, $\sigma = \sigma'$, and goes back to 2. Otherwise, it discards σ' and goes back to step 2.

The initial samples are discarded (burn-in period) until the Markov chain approaches its stationary distribution and so samples from the chain are samples from the distribution of interest. Then, the above process in repeated until the algorithm generates a given number of permutations. Recall that so far we assumed that the central permutation is the identity, $e = [123 \ldots n]$. If not, we can center the sample around σ_0 by composing each of the permutations π in the sample with σ_0, obtaining $\pi\sigma_0$.

As we have already stated, the generation of each permutation σ' is done by swapping two items in the previous permutation σ. If $p(\sigma') > p(\sigma)$ then the new permutation is accepted and the chain moves to σ'. Otherwise, the chain moves with probability $p(\sigma')/p(\sigma)$. This ratio of the probabilities can be easily computed. For both HM, where the probability is in terms of the distance, it is only necessary to check if the swapped positions where fixed points and if they are fixed points after the swap. If the distance has increased in d' then the probability of accepting the new permutation σ' is as follows:

$$exp(-\theta d')$$

For the WHM case, the probability relays on the distance decomposition vector $h(\sigma)$. If the new permutation σ' was built by swapping positions i and j from σ, then vector $h(\sigma')$ has changed in at most two positions with respect $h(\sigma)$. In case the total distance has decreased, the algorithm accepts the new permutation and the chain moves to σ'. Otherwise, the ratio of accepting σ' is follows:

$$exp(-h'_j * \theta_j - h'_i * \theta_i + h_j * \theta_j - h_i * \theta_i)$$

The computational complexity is thus $O(n)$ for the generation of each permutation. It is thus a quick algorithm. We should remark that this is an approximated algorithm.

As related works, [6] construct Markov chain algorithms for sampling from discrete exponential families conditional on a sufficient statistic.

5.3 Distances sampling

The third method we propose to generate random permutations, samples directly from the distribution. It can generate permutations only for the HM. The reason is that it is based on the fact that every permutation at the same distance from the identity under a HM model has equal probability value. This fact that not hold for the WHM case. Moreover, recall that the right invariant property implies that every permutation of n items has the same number of permutations at each distance. Assuming that the number of permutations at each possible distance d is $S_h(n, d)$, the previous fact implies that the probability of a permutation at distance d is as follows:

$$p(\sigma|d(\sigma, \sigma_0) = d) = \frac{S_h(n, d)exp(-\theta d)}{\psi(\theta)} \tag{12}$$

Note that the normalization constant $\psi(\theta) = \sum_\sigma exp(-\theta d(\sigma\sigma_0^{-1}))$ can be expressed as the sum of n terms in the following way:

$$\psi(\theta) = \sum_{d=0}^{n} S_h(n, d)exp(-\theta d) \tag{13}$$

Therefore, the process of generating a permutation from a given HM distribution can divided in three different stages as follows.

- Randomly select the distance d at which the permutation will lay using Equation (12).

- Randomly generate one permutation among those at distance d from the identity. Recall that the random generation of permutations at a prescribed Hamming distance is explained in Section 2.1.3.

- In case $\sigma_0 \neq e$ Hamming's invariance property lets us obtain a permutation centered around σ_0 as follows: $\sigma = \pi\sigma_0$ since $d = d(\pi) = d(\pi\sigma_0, \sigma_0) = d(\sigma, \sigma_0)$.

Therefore, the time complexity of the generation of each permutation using this method is $O(n)$. Summarizing, this is a fast algorithm for the generation of permutations from the HM model. Moreover, it generates samples directly form the distribution, not in an approximated way. Unfortunately, this algorithm can not generate samples from the WHM model.

6 Estimation

The parameters of a distribution given a sample of permutations are traditionally fitted via maximum likelihood estimation. In [3] it is stated that the maximum likelihood estimate of the parameters of a L-decomposable distribution can be done by iterative scaling. Unfortunately, this only includes the dispersion parameters and the exact learning of the central permutation will imply a search on the space of permutations.

In this section we address the MLE of the parameters of both HM and WHM in a different way. Recall that that HM is equal to the WHM in the case where all n dispersion parameters are equal. Therefore, we can consider in general the WHM case. For a sample of m i.i.d. permutations $\{\sigma_1, \ldots, \sigma_m\}$ the MLE for the parameters of the distribution is given by the following equation:

$$Ln \mathcal{L}(\{\sigma_1, \sigma_2, \ldots, \sigma_m\}|\sigma_0, \boldsymbol{\theta}) = \sum_{s=1}^{m} Ln \ p(\sigma_s|\sigma_0, \boldsymbol{\theta})$$

Even though the HM is a particular case of the WHM, the MLE for the parameters of the distribution are different for each distance. Therefore, we describe the model for each distance separately.

6.1 Hamming Mallows

The log-likelihood expression for the HM model is as follows:

$$Ln \prod_{i=1}^{m} p(\sigma_i) = Ln \prod_{i=1}^{m} exp(-\theta d(\sigma_i))/\psi(\theta) = \sum_{i=1}^{m} (-\theta d(\sigma_i) - Ln\,\psi(\theta)) = -m\theta\bar{d} - mLn\,\psi(\theta) \tag{14}$$

where $\bar{d} = \sum_{i=1}^{m} d(\sigma_i)/m$. By looking at Equation (14), we can see that calculating the value of σ_0 that maximizes the equation is independent of θ. Therefore the MLE estimation problem can be posed as a two step process in which first the central permutation is obtained and then the dispersion parameter for the given $\hat{\sigma}_0$.

Consensus permutation   The MLE for the consensus permutation is given by permutation that minimizes the sum of the distances to the sample. Let us find a different but equivalent formulation for the problem. Suppose M is a square matrix of $n \times n$ obtained by the sum of the permutation matrices of the m permutations in the sample. In other words, M is the frequency matrix where $M_{i,j}$ counts the number of permutations s such that $\sigma_s(i) = j$. Then, the problem of finding the permutation that minimizes the distance to the sample is equivalent to the problem of selecting one entry of M per row and column in such a way that their sum is maximum. Actually, this is equivalent to the linear assignment problem (LAP) when the objective is to maximize the sum of the assignment. The good news is that the Hungarian algorithm solves this problem in $O(n^3)$.

Dispersion parameter   Once the consensus permutation is known, the MLE for the dispersion parameter is obtained by deriving Equation (14) and equaling to zero. Their expression is then given by the next equation

$$-\bar{d} = \frac{\delta\psi(\theta)/\delta\theta}{\psi(\theta)} = \frac{\sum_{k=0}^{n} -a_k k (exp(-\theta) - 1)^{k-1} exp(-\theta)}{\sum_{k=0}^{n} a_k (exp(-\theta) - 1)^k} \tag{15}$$

where $a_k = \frac{g(n,k)}{(n-k)!k!}$. Although no closed form for θ in Equation (15) exists, root finding algorithms such as Newton-Raphson can efficiently recover θ.

As a summary, the estimation of the MLE parameters of a HM model is computationally easy.

6.2 Weighted Hamming Mallows

In this section we describe the maximum likelihood estimation process of a given sample comming from a WHM model. The log-likelihood can be expressed as follows:

$$\begin{aligned}
Ln \prod_{i=1}^{m} p(\sigma_i) =& Ln \prod_{i=1}^{m} exp(\sum_{j=1}^{n} -\theta_j h_j(\sigma_i))/\psi(\boldsymbol{\theta}) \\
=& \sum_{i=1}^{m} (\sum_{j=1}^{n} -\theta_j h_j(\sigma_i(j)) - Ln\,\psi(\boldsymbol{\theta})) \\
=& m \sum_{j=1}^{n} -\theta_j \bar{h}_j - mLn\,\psi(\boldsymbol{\theta}) \\
=& -m(\sum_{j=1}^{n} \theta_j \bar{h}_j + Ln \sum_{k=0}^{n} g(n,k)\gamma_k(T_1,\ldots,T_n))
\end{aligned} \tag{16}$$

where $T_i = (exp(-\theta_i) - 1)$ and $\bar{h}_j = \sum_{i=1}^{m} h_j(\sigma_i)/m$. Given the σ_0 that maximizes the likelihood, the dispersion parameters are given by the derivative equals 0 of the likelihood for θ_i

$$-h_i = \frac{\sum_{k=0}^{n} g(n,k)\frac{\delta\gamma_k}{\delta\theta_i}(T_1,\ldots,T_n)}{\sum_{k=0}^{n} g(n,k)\gamma_k(T_1,\ldots,T_n)}$$

where by $\frac{\delta\gamma_k}{\delta\theta_i}(T_1,\ldots,T_n)$ we mean the derivative of the elementary symmetric polynomial of $\gamma_k(T_1,\ldots,T_n)$ with respect to θ_i. The efficient computation for both elementary symmetric polynomial $\gamma_k(T_1,\ldots,T_n)$ and its derivatives is given in Section 2.1.4.

7 Proofs

Proofs for the derivatives of the ESP in Section 2.1.4.

Proof. We start by explicitly giving the expansion of the ESP on four variables of the form $X_i = (exp(-\theta_i) - 1)$, as we will find in the normalization constant.

$$\gamma_1 = \begin{cases} \gamma_1^1 = \begin{cases} (exp(-\theta_1) - 1) \\ (exp(-\theta_2) - 1) \\ \bar{\gamma}_1^1 = \begin{cases} (exp(-\theta_3) - 1) \\ (exp(-\theta_4) - 1) \end{cases} \end{cases} \end{cases} \qquad \gamma_2 = \begin{cases} \gamma_2^1 = \begin{cases} (exp(-\theta_1) - 1)(exp(-\theta_2) - 1) \\ (exp(-\theta_1) - 1)(exp(-\theta_3) - 1) \\ (exp(-\theta_1) - 1)(exp(-\theta_4) - 1) \\ \bar{\gamma}_2^1 = \begin{cases} (exp(-\theta_2) - 1)(exp(-\theta_3) - 1) \\ (exp(-\theta_2) - 1)(exp(-\theta_4) - 1) \\ (exp(-\theta_3) - 1)(exp(-\theta_4) - 1) \end{cases} \end{cases} \end{cases}$$

$$\gamma_3 = \begin{cases} \gamma_3^1 = \begin{cases} (exp(-\theta_1) - 1)(exp(-\theta_2) - 1)(exp(-\theta_3) - 1) \\ (exp(-\theta_1) - 1)(exp(-\theta_2) - 1)(exp(-\theta_4) - 1) \\ (exp(-\theta_1) - 1)(exp(-\theta_3) - 1)(exp(-\theta_4) - 1) \\ \bar{\gamma}_3^1 = \begin{cases} (exp(-\theta_2) - 1)(exp(-\theta_3) - 1)(exp(-\theta_4) - 1) \end{cases} \end{cases} \end{cases}$$

$$\gamma_4 = \begin{cases} \gamma_4^1 = \begin{cases} (exp(-\theta_1) - 1)(exp(-\theta_2) - 1)(exp(-\theta_3) - 1)(exp(-\theta_4) - 1) \\ \bar{\gamma}_4^1 = \begin{cases} 0 \end{cases} \end{cases}$$

We will now consider the expansion of γ_k

$$\gamma_k = \sum_{r=0}^{k}(-1)^{k-r}\alpha_{k,r}^n exp(\sum_{j \in A} -\theta_j)\text{where } A \in \{1, \ldots, n\} \wedge |A| = r \tag{17}$$

Recall that γ_k can be divided (as in equation (3)) into two different subset of sums, those that include $(exp(-\theta_i) - 1)$ and those that do not include $(exp(-\theta_i) - 1)$. The expansion of the latter $\bar{\gamma}_k^i$ is

$$\bar{\gamma}_k^i = \sum_{r=0}^{k}(-1)^{k-r}\beta_{k,r}^n exp(\sum_{j \in A} -\theta_j)\text{where } A \in \{1, \ldots, n\} \wedge |A| = r \wedge i \notin A \tag{18}$$

Note that the coefficients in the expansions of γ_k and $\bar{\gamma}_k^i$ are different -namely $\alpha_{k,r}^n$ and $\beta_{k,r}^n$ they are somehow related as we will later see. Let us show how $\alpha_{k,r}^n$ are obtained. We have to count the number of addends in which the same group of r different variables will appear.

- The number of addends i.e. number of lines in the previous example, is $\binom{n}{k}$.

- There are k terms in each addend. Once an addend is expanded by multiplying the terms there will be $\binom{k}{r}$ terms of degree r per addend, i.e. like $exp(\sum_{j \in A} -\theta_j)$ for $|A| = r$.

- There are $\binom{n}{r}$ groups of r items.

Therefore,

$$\alpha_{k,r}^n = \frac{\binom{k}{r}\binom{n}{k}}{\binom{n}{r}} = \frac{(n-r)!}{(k-r)!(n-k)!}$$

With the same reasoning we obtain the coefficients $\beta_{k,r}^n$

$$\beta_{k,r}^n = \frac{\binom{k}{r}\binom{n-1}{k}}{\binom{n-1}{r}} = \frac{(n-1-r)!}{(k-r)!(n-1-k)!}$$

The first derivative with respect to θ_i of the expansion in Equation (17) cancels every term that does not include θ_i.

$$\frac{\delta\gamma_k}{\delta\theta_i} = (-1) * \sum_{r=1}^{k}(-1)^{k-r}\alpha_{k,r}^n exp(\sum_{j \in A} -\theta_j) \text{ where } A \in \{1, \ldots, n\} \wedge |A| = r \wedge i \in A$$

Since θ_i appears in every term, it is possible to take out this common factor, resulting

$$\frac{\delta\gamma_k}{\delta\theta_i} = (-exp(\theta_i)) * \sum_{r=0}^{k-1}(-1)^{(k-1)-r}\alpha_{k,r+1}^n exp(\sum_{j \in A} -\theta_j) \text{ where } A \in \{1, \ldots, n\} \wedge |A| = r \wedge i \notin A$$

15

Let us now focus on the coefficient $\alpha_{k,r+1}^n$:

$$\alpha_{k,r+1}^n = \frac{(n-r-1)!}{(k-r-1)!(n-k)!} = \beta_{k-1,r}^n = \frac{(n-1-r)!}{(k-1-r)!(n-1-k+1)!}$$

So the above expression can be rewritten as

$$\frac{\delta \gamma_k}{\delta \theta_i} = (-exp(\theta_i)) * \sum_{r=0}^{k-1} (-1)^{(k-1)-r} \beta_{k-1,r}^n exp(\sum_{j \in A} -\theta_j) \text{ where } A \in \{1, \ldots, n\} \wedge |A| = r \wedge i \notin A$$

Finally, note that part of the above expression is a particular case of Equation (18), so we can equivalently write it as follows, finishing the proof.

$$\frac{\delta \gamma_k}{\delta \theta_i} = -exp(-\theta_i)\bar{\gamma}_{k-1}^i$$

$\square$

The computational complexity of $\frac{\delta \gamma_k}{\delta \theta_i}$ given $\bar{\gamma}_k^i \ \forall k, i$ is $O(n^2)$. Let us now give an example of the expansion of the ESP γ_3 for $n = 4$.

$$\begin{aligned}
\gamma_3 =& exp(-\theta_1 - \theta_2 - \theta_3) + exp(-\theta_1 - \theta_2 - \theta_4) + exp(-\theta_1 - \theta_3 - \theta_4) + exp(-\theta_2 - \theta_3 - \theta_4) \\
& - 2exp(-\theta_1 - \theta_2) - 2exp(-\theta_1 - \theta_3) - 2exp(-\theta_1 - \theta_4) - 2exp(-\theta_2 - \theta_3) - 2exp(-\theta_2 - \theta_4) - 2exp(-\theta_3 - \theta_4) \\
& + 3exp(-\theta_1) + 3exp(-\theta_2) + 3exp(-\theta_3) + 3exp(-\theta_4) \\
& - 4
\end{aligned}$$

Here $\alpha_{3,3}^4 = 1, \alpha_{3,2}^4 = 2, \alpha_{3,1}^4 = 3, \alpha_{3,0}^4 = 4$. The expansion of γ_2 for $n = 4$ is as follows

$$\begin{aligned}
\gamma_2 =& exp(-\theta_1 - \theta_2) + exp(-\theta_1 - \theta_3) + exp(-\theta_1 - \theta_4) + exp(-\theta_2 - \theta_3) + exp(-\theta_2 - \theta_4) + exp(-\theta_3 - \theta_4) \\
& - 3exp(-\theta_1) - 3exp(-\theta_2) - 3exp(-\theta_3) - 3exp(-\theta_4) \\
& + 6
\end{aligned}$$

On the other hand, $\alpha_{2,2}^4 = 1, \alpha_{2,1}^4 = 3, \alpha_{2,0}^4 = 6$.

$$\begin{aligned}
\gamma_2^1 =& exp(-\theta_1 - \theta_2) + exp(-\theta_1 - \theta_3) + exp(-\theta_1 - \theta_4) \\
& - 3exp(-\theta_1) - exp(-\theta_2) - exp(-\theta_3) - exp(-\theta_4) \\
& + 3
\end{aligned}$$

$$\begin{aligned}
\bar{\gamma}_2^1 =& exp(-\theta_2 - \theta_3) + exp(-\theta_2 - \theta_4) + exp(-\theta_3 - \theta_4) \\
& - 2exp(-\theta_2) - 2exp(-\theta_3) - 2exp(-\theta_4) \\
& + 3
\end{aligned}$$

Here $\beta_{2,2}^4 = 1, \beta_{2,1}^4 = 2, \beta_{2,0}^4 = 3$.

$$\begin{aligned}
\frac{\delta \gamma_3}{\delta \theta_1} =& -exp(-\theta_1 - \theta_2 - \theta_3) - exp(-\theta_1 - \theta_2 - \theta_4) - exp(-\theta_1 - \theta_3 - \theta_4) \\
& + 2exp(-\theta_1 - \theta_2) + 2exp(-\theta_1 - \theta_3) + 2exp(-\theta_1 - \theta_4) \\
& - 3exp(-\theta_1)
\end{aligned}$$

The derivative $\frac{\delta \gamma_3}{\delta \theta_1}$ can be given as a function of $\bar{\gamma}_2^1$ as follows:

$$\begin{aligned}
\frac{\delta \gamma_3}{\delta \theta_1} =& -exp(-\theta_1)[exp(-\theta_2 - \theta_3) + exp(-\theta_2 - \theta_4) + exp(-\theta_3 - \theta_4) \\
& -2exp(-\theta_2) - 2exp(-\theta_3) - 2exp(-\theta_4) \\
& +3] \\
=& -exp(-\theta_1)\bar{\gamma}_2^1
\end{aligned}$$

$$\bar{\gamma}_0^{i,j} = 1$$
$$\bar{\gamma}_1^{i,j} = \sum_{r \neq i,j} (exp(-\theta_r) - 1)$$
$$\bar{\gamma}_k^{i,j} = \gamma_k - \bar{\gamma}_k^i - \gamma_k^j + \gamma_k^{i,j}$$
$$\gamma_k^{i,j} = (exp(-\theta_i) - 1) * (exp(-\theta_j) - 1)\bar{\gamma}_{k-2}^{i,j}$$

7.1 more things, ESP

This can lead to an efficient sampling algorithm

$$\bar{\gamma}_{k+1}^i = \bar{\gamma}_{k+1}^{ij}(exp(-\theta_j) - 1) + \bar{\gamma}_{k+1}^{ij}$$

8 Conjetures

sufficient statistic matrix iff l-decomposable ?

Complexity. If the distance between two permutations of n items ranges between 0 and max_dist then the complexity of generating a random permutation at distance d, $0 \le d \le max\_dist$ is $O(max\_dist)$.

References

[1] Frank B Baker and Michael R Harwell. Computing Elementary Symmetric Functions and Their Derivatives: A Didactic. *Applied Psychological Measurement*, 20(2):169–192, 1996.

[2] Douglas Edward Critchlow, Michael A Fligner, and Joseph S Verducci. Probability Models on Rankings. *Journal of Mathematical Psychology*, 35:294–318, 1991.

[3] Vill Csiszár. On L-decomposability of random orderings. *Journal of Mathematical Psychology*, 53(4):294–297, 2009.

[4] Angela D'Elia and Domenico Piccolo. A mixture model for preferences data analysis. *Computational Statistics & Data Analysis*, 49(3):917–934, 2005.

[5] Persi Diaconis. *Group representations in probability and statistics*. Institute of Matematical Statistics, 1988.

[6] Persi Diaconis and Bernd Sturmfels. Algebraic Algorithms for Sampling from Conditional Distributions. *The Annals of Statistics*, 26(1), 1998.

[7] Michael A Fligner and Joseph S Verducci. Distance based ranking models. *Journal of the Royal Statistical Society*, 48(3):359–369, 1986.

[8] Alexander Gnedin and Grigori Olshanski. The two-sided infinite extension of the Mallows model for random permutations. *Advances in Applied Mathematics*, 48(5):615–639, 2012.

[9] Duncan Luce R. *Individual Choice Behavior*. Wiley, New York, 1959.

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

[11] Yi Mao and Guy Lebanon. Non-Parametric Modeling of Partially Ranked Data. *Journal of Machine Learning Research*, 9:2401–2429, 2008.

[12] Marina Meila and Le Bao. Estimation and Clustering with Infinite Rankings. In *Uncertainty in Artificial Intelligence (UAI)*, pages 393–402, Corvallis, Oregon, 2008. AUAI Press.

[13] Marina Meila and Harr Chen. Dirichlet Process Mixtures of Generalized Mallows Models. In *Uncertainty in Artificial Intelligence (UAI)*, pages 285–294, 2010.

[14] Thomas Brendan Murphy and Donal Martin. Mixtures of distance-based models for ranking data. *Computational Statistics & Data Analysis*, 41(34):645–655, 2003.

[15] R L Plackett. The Analysis of Permutations. *Journal of the Royal Statistical Society*, 24(10):193–202, 1975.