Inversion Models beyond sufficient statistics

Marina Meilă
University of Washington
Seattle, WA 98195
mmp@stat.washington.edu

Christopher Meek
Microsoft Research
Redmond, WA 98052
meek@microsoft.com

Abstract

Can we do exact and tractable inferences in Mallows-like models for incomplete data? We show here that the answer is *yes* for the most general form Mallows-type model and a large class of partial orders known as *partial rankings*. Top-t rankings and ratings are special cases of partial rankings.

1 Inversion models and their sufficient statistics

Mallows-type models have recently received much interest in the literature on modeling rankings, and a comensurably large number of applications. This type of models are exponential families, where the statistics are inversion counts, so that the probability of a permutation π is penalized exponentially in the number of inversions w.r.t a *central permutation* π_0.

Models in this class have been studied, among others, by (Fligner and Verducci, 1986),(Meilă et al., 2007), and recently by (Meek and Meilă, 2014). The present work will focus on the latest model, the RIM , which was shown to include the Mallows and Generalized Mallows models as subclasses. Our results specialize naturally to these sub-classes.

For any inversion-counting model, it is natural to represent a permutation π over the set of items $\mathcal{E} = \{e_1, \dots e_n\}$ by its *discrepancy matrix* $Q(\pi)$ given by

$$Q(\pi) \;=\; [Q_{ee'}]_{e,e'=1}^{n}, \qquad Q_{ee} = 0,\ Q_{ee'} = 1 \text{ if } e \prec_\pi e' \text{ and 0 otherwise, for all } e, e' \in E,\ e \neq e'$$

where $e \prec_\pi e'$ means "e precedes e' in π". To note that it is sufficient to know the upper triangle $e < e'$ of the discrepancy Q, since $Q_{ee'} = 1 - Q_{e'e}$ and $Q_{ee} = 0$.

A RIM $\tau(\vec{\theta})$ is a distribution over the permutations of a set $\mathcal{E} = \{e_1, \dots, e_n\}$, which has a structure τ and a set of parameters $\theta_{1:n-1} \geq 0$; τ is a binary tree with n leaves, each associated with a distinct element of $\mathcal{E}$. We denote the set of internal vertices of the binary tree by $\mathcal{I}$ and each internal vertex is represented as a triple $i = (i_L, i_R, \theta_i)$ where i_L (i_R) is the left (right) subtree, and θ_i controls the number of inversions when merging the subsequences generated from each of the subtrees. Traversing the tree τ *in preorder*, with the left child preceding the right child induces a permutation on $\mathcal{E}$ called the *reference permutation* of the RIM which we denote as π_τ.

The *number of inversions at (internal) vertex* i of $\tau(\vec{\theta})$ for test permutation π is $v_i(\pi, \pi_\tau) = \sum_{l \in \mathcal{L}_i} \sum_{r \in \mathcal{R}_i} Q_{lr}(\pi, \pi_\tau)$ where $\mathcal{L}_i$ ($\mathcal{R}_i$) is the subset of $\mathcal{E}$ under i_L (i_R), the left (right) subtree of internal vertex i. The likelihood of a permutation π with respect to RIM $\tau(\vec{\theta})$ is as follows:

$$P(\pi|\tau(\vec{\theta})) \propto \prod_{i \in \mathcal{I}} \exp(-\theta_i v_i(\pi, \pi_\tau))/Z_{L_i,R_i}(\theta_i) \quad \text{with } Z_{n,m}(\theta) = \frac{(\theta)_{n+m}}{(\theta)_n (\theta)_m} \text{ and } (\theta)_n = \prod_{k=1}^{n} \frac{1 - e^{-k\theta}}{1 - e^{-\theta}}$$

It has been shown that, if the data consists of a set of N complete rankings, sampled i.i.d, then the data discrepancy $Q = \sum_{\pi \in \mathcal{D}} Q(\pi)$ plays the role of sufficient statistics for estimating both the structure τ and the parameters θ at the internal nodes. (Meek and Meilă, 2014) also present algorithms for ML estimation of $\tau, \vec{\theta}$.

Our aim here is to examine the more realistic scenario where preference data is collected from independent rankers (e.g. users) in the form of partial orders. In this scenario, unfortunately, there typically will be no sufficient statistics to allow us to compress the data[1]. The question we ask is, what else is lost when the observations are not complete? To what extent does the (ML estimation) problem become harder for partial observations, and what algorithms can perform it?

2 The marginal likelihood of a partial ranking

Here we provide very encouraging answers for a special case of incomplete observations called *partial rankings*. A *partial ranking* is a partial order on $\mathcal{E}$ defined by a partion of $\mathcal{E}$ into disjoint subsets $E_1, \ldots E_K$, and a total ordering of the subsets. We denote a partial ranking by $\sigma = (E_1|E_2|\ldots|E_K)$, meaning that all elements of E_1 preceed the elements of E_2, the elements of E_2 in turn preceed all the elements of E_3, etc. If $n_1 = |E_1|, \ldots n_K = |E_k|$, with $n = \sum_k n_k$, then we call σ a partial ranking of *shape* $(n_1, \ldots n_K)$ with K *grades*.

Mallows models over partial rankings were introduced by (Mao and Lebanon, 2008). Rating all items of a set, by e.g *, **, … ***** produces a partial ranking with $K = 5$. An important example are the rankings of shape $(\underbrace{1, 1, \ldots 1}_{\times t}, n - t)$, which correspond to observing the top t ranks of π.

In (Huang et al., 2012) it was shown that, in a *Riffle Independence (RI)* model (Huang and Guestrin, 2012), the likelihood of any partial ranking factors according to the model[2]. As RIM's are a special case of RI model, with a compact parametrization, it follows that the likelihood of any partial ranking is a product of factors, each corresponding to an internal node in the RIM.

However, for a RI model, the only known algorithm for calculating the likelihood involves enumerating over all possible interleavings at a node, which is exponential in the size of the smallest child of the node. Hence, this algorithm is practical only for Mallows models and very imbalanced trees, or for small n. In contrast, the first result we present is that for the compactly parametrized RIM, computing the likelihood of a partial ranking is tractable.

Theorem 1 *Let $\sigma = (E_1|E_2|\ldots|E_K)$ be a partial ranking, $\tau(\vec{\theta})$ a RIM model. Choose an internal node i of τ, and let $\mathcal{L}, \mathcal{R}$ be sets of items in the left and right subtrees of i (for simplicity, we drop i from the remainder of this definition). Denote $L_k = E_k \cap \mathcal{L}$, $R_k = E_k \cap \mathcal{R}$, $l_k = |L_k|, r_k = |R_k|$, $\bar{l} = (l_1, \ldots l_K)$, $\bar{r} = (r_1, \ldots r_K)$ and θ be the parameter at node i. Define*

$$g(\bar{l}, \bar{r}, \theta) \;=\; \frac{Z_{l_1,r_1}(\theta) Z_{l_2,r_2}(\theta) \ldots Z_{l_K,r_K}(\theta) \theta^{(l_2+\ldots l_K)r_1 + (l_3+\ldots l_K)r_2 + \ldots + (l_2+\ldots l_K)r_{K-1}}}{Z_{|\mathcal{L}|,|\mathcal{R}|}(\theta)} \tag{1}$$

Then the likelihood of σ in the mode $\tau(\vec{\theta})$ is

$$P_{\tau(\vec{\theta})}(\sigma) \;=\; \prod_{i \in \mathcal{I}} g(\bar{l}_i, \bar{r}_i, \theta_i) \tag{2}$$

It is instructive to compare expression (1) with the corresponding expression for a complete permutation $\theta^v / Z_{l,r}(\theta)$. One sees that missing information traslates into up to $K - 1$ additional Z factors in the numerator.

It can thus appear that computing the marginal of a partial ranking is K times more expensive than that of a complete ranking, as for the latter, at each node one computes a single Z. However, a more careful look shows that this is not so.

Theorem 2 *The number of $Z_{l,r}$ evaluations for any partial ranking is no larger than $2(n-1)$, i.e it is no larger than twice the number of Z evaluations for a complete permutation.*

[1] To note that for very small sample sizes, or for Mallows "deletion models", the sufficient statistics are not always taking less storage than the data.

[2] This property is called *complete decomposablility* and partial rankings are the only completely decomposable partial orders.

We sketch the proof of this result. First, note that, $Z_{l,0} = Z_{0,r} = 1$ for all r, l. Hence, every time one of L_k, R_k are empty, the correspoding factor in (1) becomes equal to 1. If a set E_k has n_k elements, the total number of nodes where both L_k, R_k are non-empty is $n_k - 1$. The total number of non-trivial factors in the numerators of all $g(l^i, r^i, \theta_i)$ is no larger than $\sum_k (n_k - 1) = n - K < n - 1$. Hence, there are no more extra Z factors than nodes in τ. We remark that this bound is not tight, and in most cases it is a gross exaggeration of the actual computation overhead. For example, let us calculate explicitly the (marginal) probability of a top-t ranking $\sigma = (e_1|e_2|\ldots|e_t|\mathcal{E} \setminus \{e_1, \ldots e_t\})$.

Let l_t, r_t, $l_t + r_t = t$ denote the number of top-t items in the left, respectively right subsets of a node. Let $0 \leq \tilde{v}' \leq rl$ denote the number of inversions among the top-t items. For example, the pattern $l|r|r|l|r|l$ has $\tilde{v}' = 2 + 2 + 1 = 5$ inversions. Now denote $\tilde{v} = \tilde{v}' + r_t(L - l_t)$; for any π consistent with top-t ranking σ, the number of inversions at at node is greater or equal to $\tilde{v}$. Therefore,

$$g(\bar{l}, \bar{r}, \theta) \;=\; \frac{\theta^{\tilde{v}} Z_{L-l_t, R-r_t}(\theta)}{Z_{L,R}(\theta)} \qquad (3)$$

In total, there will be no more than t additional Z's to compute, which can be much less that $n - 1$, the number of Z'z for complete permutations.

Moreover, as we will show in the extended paper, once the denominator is computed, the overhead for obtaining all the factors in the numerator is minor in comparison.

3 Model estimation from partial rankings

There are two aspects to this problem: *parameter estimation*, i.e estimatiion of $\theta_{1:n-1}$ given data and a structure τ, and *structure estimation* where τ is estimated from data, with the parameter estimation as an inner loop. For complete π's parameter estimation is "easy" (convex, univariate) while structure estimation is either provably NP-hard or of suspected to be so. In this context, one should take tractability to mean at most $Npoly(n)$ times more expensive than estimation from complete data. The state of the art for estimating RI models from partial rankings is the EM algorithm proposed by Huang et al. (2012). Such an algorithm can also be used for a RIM. In the E step, the partial rankings are "completed" to complete rankings. Since there can be an exponential number of such completions, the exact E step is intractable, and the actual algorithm samples from the conditional distribution given the observations. For the RIM , and complete permutations, (Meek and Meilă, 2014) proposed a search algorithm, SASEARCH, that consists of alternating steps, all tractable, of sampling and optimizing over τ given a π_τ.

For partial rankings, we show that[3]: (1) the parameter estimation problem is convex, and computing the gradient is $\mathcal{O}(Npoly(n))$; (2) all the tractable steps of SASEARCH can be performed exactly and tractably for partial rankings; (3) in this case SASEARCH is directly maximizing the likelihood (i.e. it's NOT and EM algorithm); (4) if one was to use an EM algorithm then (obviously) the M step is (equivalent to) a ML estimation from complete data. Any complete data ML estimation algorithm, in particular SASEARCH can be used as an M step. Hence, the interest is in the E step, which is specific to the kind of missingness of the data at hand. Thus, (5), we propose a different E step, both exact and efficient. Instead of completing the *permutations* as () propose, we complete the *sufficient statistics* Q. We present the idea on an example in Figure 1. Our proposed algorithm, FILLQ, uses the subroutine INTERLEAVPROB, derived by us and part of the extended version of (Meek and Meilă, 2014). Given an interleaving $(\mathcal{L}, \mathcal{R}, \theta)$, algorithm INTERLEAVPROB computes the probabilities $Q_{lr} = Pr[l \prec r$ for all $l \in \mathcal{L}, r \in \mathcal{R}$. Once the matrix $Q(\sigma)$ has been completed for one observation, it is added to the sufficient statistics of the previous observations. Thus the output of the E step is a matrix of sufficient statistics, which represent the correct marginals given the partial rankings observed. On this matrix Q, a maximization step can proceed as usual. We stress that we put no restriction on the kind of partial rankings that can be observed, nor do we constrain one observation to have similar structure to another[4]. For example, the observations could each have a different number of grades K, or they can be top-t rankings with different lengths t, or they could include top-$t$and and bottom-t rankings. The whole E step will not incur more overhead than an (amortized) factor of N.

[3] Details in extended version.

[4] But, whenever two or more observation share some structure, with simple caching we can save computations.

$\sigma = (cf|d|abe)$
Original $Q(\sigma)$

	a	b	c	d	e	f
a	–	?	0	0	?	0
b		–	0	0	?	0
c			–	1	1	?
d				–	1	0
e					–	0
f						–

$\tau(\vec{\theta})$
$(((a, b, \theta_3)(c, d, \theta_4), \theta_2), (e, f, \theta_5), \theta_1)$

=

node $(abcd, ef, \theta_1)$
$L_1 = \{c\}, R_1 = \{f\}$ get Q_{cf}=InterleavProb$(1, 1, \theta_1)$
$L_2 = \{d\}, R_2 = \emptyset$ nothing to compute
$L_3 = \{a, b\}, R_3 = \{e\}$ get Q_{ae}, Q_{be}=InterleavProb$(2, 1, \theta_1)$
node (ab, cd, θ_2), $\sigma = (c|d|ab)$
$L_1 = \emptyset, R_1 = \{c\}; L_2 = \emptyset, R_2 = \{d\}; L_3 = \{a, b\}, R_3 = \emptyset$ nothing to compute
node (a, b, θ_3), $E = (ab)$
$L = \{a\}, R = \{b\}$ get Q_{ab}=InterleavProb$(1, 1, \theta_3)$
node (e, f, θ_4), $E = (f|e)$
$L_1 = \emptyset, R_1 = \{f\}; L_2 = \{e\}, R_2 = \emptyset$ nothing to compute

Figure 1: Algorithm FillQ$(\sigma, \tau(\vec{\theta}))$ computes the expected inversion probabilities given a model $\tau(\vec{\theta})$ and a partial ranking σ. Left: the original $Q(\sigma)$, with some entries observed (e.g $Q_{cd} = Q_{fd} = 1$), and the others to be filled in with conditional probabilities given the observations. Right: the algorithm proceeds recursively top-down; each time when $L_k^i, R_k^i \neq \emptyset$ some $Q_{ee'}$ are calculated based on the θ_i of the current node i. FillQ outputs a sufficient statistics matrix Q, with $Q_{ee'} = Pr[e \prec e'|\sigma]$.

4 Contributions

We consider the RIM, a superclass of Mallows and GMM and the most general inversion counting model that has tractable Z known to date. We consider partial rankings, a class of partial observations that include top-$t$rankings, ratings and other common observation models. We obtain

1. Exact formulas and polynomial algorithms for computing the marginal probability of a partial ranking in a RIM $\tau(\vec{\theta})$. We show that this is no more than twice as expensive as the likelihood of a complete ranking (and often much less). This generalizes and extends previous work of (Mao and Lebanon, 2008) that computed the same marginals for single parameter, unstructured, Mallows models.
2. Exact recursive algorithm FillQ for computing the pairwise marginals $Q_{ee'}$ conditioned on a partial ranking. This algorithm is faster than the corresponding algorithm for a complete ranking, as it should be, given that only a subset of pairs must be considered. This algorithm differs significantly from the algorithm presented in Huang et al. (2012) in that we do not enumerate and renormalize explicitly, but use polynomial recursions to obtain our result.
3. We show that the ML parameter estimation at each node is a convex univariate problem, and give an exact polynomial algorithm for computing the gradient (not included). We show that with this, structure search can proceed as if the rankings were complete.
4. We introduce a new E step for the model estimation, that completes the *sufficient statistics* directly, instead of the data, leading to faster and more efficient overall estimation.

Using sound statistical techniques, exact and efficient algorithms are is desirable in the analysis of rank data. So is being able to analyze realistic input data, marred by incompleteness or uncertainty. This paper shows that a researcher need not compromise in either aspect, as we develop the methodology to make exact inferences with partial rankings, and let one do it almost as efficient as for full rankings.

References

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

Huang, J. and Guestrin, C. (2012). Uncovering the riffled independence structure of ranked data. *Electronic Journal of Statistics*, 6:199–230.

Huang, J., Kapoor, A., and Guestrin, C. (2012). Riffled independence for efficient inference with partial rankings. *Journal of Artificial Intelligence Research*, 44:491–532.

Mao, Y. and Lebanon, G. (2008). Non-parametric modelling of partially ranked data. *Journal of Machine Learning Research*, 9:2401–2429.

Meek, C. and Meilă, M. (2014). Recursive inversion models for permutations. In Cortes, C. and Lawrence, N., editors, *Advances in Neural Information Processing Systems*. MIT Press.

Meilă, M., Phadnis, K., Patterson, A., and Bilme s, J. (2007). Consensus ranking under the exponential model. In Parr, R. and Van den Gaag, L., editors, *Proceedings of the 23rd Conference on Uncertainty in AI*, volume 23.