Structure Discovery from Partial Rankings

Jonathan Huang
Carnegie Mellon University
jch1@cs.cmu.edu

Ashish Kapoor
Microsoft Research
akapoor@microsoft.com

Abstract

Aggregating and statistical reasoning with ranked data are tasks that arise in a number of applications from analyzing political elections to modeling user preferences over a set of items. Representing distributions over rankings, however, can be daunting due to the fact that the number of rankings of n items scales factorially. Moreover, it is crucial for probabilistic models over rankings to be able to handle partially ranked data since real world data more often consists of partial rankings rather than full. In this work, we study a class of models over rankings called hierarchical riffle independent models, which can be thought of as being analogous to graphical models but more appropriate for ranked data. We show in particular that Bayesian conditioning based on *top-k* partial ranking evidence can be performed efficiently in these models, and apply our algorithms to estimate the structure of a riffle independent model from top-k rankings.

1 Introduction

The need to aggregate and reason over large collections of rankings comes up in a variety of settings in which one must collect user preferences and judgements in order to make predictions. Rankings arise, for example, in analyzing preferences for movies [8], for elections in various political organization [5], and in using the 'wisdom of the crowds' phenomenon to piece together recalled order information [9].

Representing uncertainty over rankings is challenging, since there are $n!$ possibilities, and typical factorized representations, such as graphical models, cannot efficiently capture the mutual exclusivity constraints that are associated with permutations. In recent papers, [5, 6] introduced a tractable, decomposable family of distributions over rankings called hierarchical riffled independent models.

In this work, we model *partially ranked* data using hierarchical riffle independent models. In particular, we propose a principled approach for modeling datasets in which users rank only their k most favorite items. Using our methods, we are able to estimate the hierarchical structure of the item set from top-k rankings (like graphical model structure learning from incomplete data) as well as to make probabilistic predictions on test data. Our contributions are as follows:

- We show an efficient algorithm for Bayesian conditioning on rankings when the prior distribution is riffle independent and the evidence is in the form of a top-k ranking.
- Using our conditioning method, we are able to perform structure learning for riffle independent hierarchies using partially ranked data.

2 Riffled independence for rankings

In this paper we will be concerned with distributions over the rankings of a finite item set indexed by the set $\{1, \ldots, n\}$. A ranking is a joint assignment of items to ranks, $\sigma = (\sigma(1), \sigma(2), \ldots, \sigma(n))$, where $\sigma(i)$ denotes the rank of item i. We will refer to the set of all rankings of a set of n items as S_n. Assuming that two items are not mapped to the same rank (a condition we refer to as *mutual exclusivity*), there are $n!$ possible rankings of the item set, and consequently, it is not tractable to estimate or even explicitly represent arbitrary distributions over rankings for nontrivial n.

A popular technique for dealing with distributions over a large number of variables is to assume independence structure [7] often in the form of, for example, a naive Bayes assumption, or more generally a graphical model. Due to mutual exclusivity however, it is not obvious how one should exploit independence for rankings since there is a dependency between the ranks of every pair of items.

Generalized independence relations. *Riffled independence*, introduced by Huang and Guestrin in [5] is a generalized form of independence assumptions for permutations. Riffled independence assumptions are typically more suitable for ranked data while retaining many of the computational advantages of assuming ordinary independence. In this section, we provide a brief introduction.

We now assume that the item set X is the union of two disjoint sets A and B, which can be, without loss of generality, $\{1, \ldots, p\}$ and $\{p+1, \ldots, n\}$, respectively. For example if X is the set of candidates in a political election, A might be the set of conservative candidates, and B the set of liberal candidates. We say that the set A is *riffle independent* of B (with respect to a distribution h) if a full ranking, σ, of X can be generated via the following procedure: first draw a ranking of items in A from some distribution f over S_p; independently draw a ranking of items in B from some distribution g over S_{n-p}; and finally (in order to circumvent mutual exclusivity), independently draw an interleaving of the two sets from a distribution m defined over all possible interleavings of A and B. The probability of σ can therefore be written as the product of three terms:

$$h(\sigma) = m(\tau_{A,B}(\sigma)) \cdot f(\phi_A(\sigma)) \cdot g(\phi_B(\sigma)),$$

where m is a distribution over possible interleavings of the two sets, A and B (called the *interleaving distribution*), and f and g are distributions on rankings over just A and just B respectively (called *relative ranking distributions*). For example, if a_1, a_2 are conservative candidates and b_1, b_2 are liberal, a ranking of $X = \{a_1, a_2, b_1, b_2\}$ can be obtained by first choosing $\sigma(a_1) < \sigma(a_2)$ (a_1 preferred to a_2), then choosing $\sigma(b_2) < \sigma(b_1)$, then interleaving to obtain $\sigma(a_1) < \sigma(b_2) < \sigma(a_2) < \sigma(b_1)$.

Hierarchical decompositions. Item sets rarely partition neatly into two clusters. Instead, it is often natural to consider hierarchical decompositions of itemsets into nested collections of partitions (much like hierarchical clustering). For example, A might be riffle independent of B, but A itself might be partitioned into subsets C and D that can be thought of as *marginally* riffle independent.

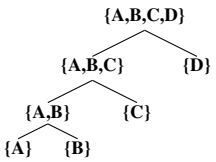

Figure 1: An example of a hierarchical structure in which the item set is decomposed into singletons

Like Bayes nets, these hierarchies represent families of distributions obeying a certain set of (riffled) independence constraints and can be parameterized locally. To draw from such a model, one generates full rankings recursively starting by drawing rankings of the leaf sets, then working up the tree, sequentially interleaving rankings until reaching the root. By decomposing huge distributions over rankings into tractable pieces (like Bayes nets have done for other distributions), these hierarchical models allow for better interpretability as well as efficient probabilistic representation and inference, as we will discuss further below.

3 Exploiting riffled independence for efficient inference

Given some representation of a distribution, we would like to perform probabilistic inference operations. We might want to answer simple questions such as "what rank is liberal candidate 1 most likely to be in?", or more complicated questions such as "given that a voter ranked candidate 4 in first place and candidate 10 in second place, what candidate is she most likely to place in last place?" The answer to both questions and other similar ones lies in Bayes rule, in which a posterior distribution is obtained by multiplying a prior and likelihood, then normalizing. Often the normalization constant itself is of interest and while its computation is conceptually simple, it is often computationally hard unless one can exploit the structure of the underlying representation in a way such that summing over probabilities is efficient. For example, to compute the marginal probability of candidate i being in rank j, one multiplies the joint distribution by the indicator $\mathbb{1}[\sigma(i) = j]$ and sums over the resulting function to obtain the desired probability. In this section, we consider performing conditioning operations on a distribution in which subsets A and B are riffle independent.

For simplicity, we begin by considering the simple problem of conditioning on $\sigma(i) = j$ (item i is ranked j) when the ranks of sets A and B are *truly* independent of each other. In this fully independent case,

$$h(\sigma) = h(\sigma(A)) \cdot h(\sigma(B)) = h(\sigma(1), \ldots, \sigma(p)) \cdot h(\sigma(p+1), \ldots, \sigma(n)),$$

and assuming that $h(\sigma(i) = j) > 0$, then instead of operating on the full distribution h, one can instead condition only one of the factors, $h(\sigma_A)$ or $h(\sigma_B)$ depending on whether $i \in A$ or $i \in B$, leading to a more efficient Bayesian update.

Conditioning on top-k data. Can we condition without updating the entire distribution when A and B are only *riffle independent*? Unfortunately, it can be shown that it is not possible to efficiently

condition on arbitrary (item, rank) associations such as $\sigma(i) = j$. We can show, however, that something surprising happens when conditioning on $\sigma(i) = 1$ (item i is ranked in first place).

In this special case, it turns out that the indicator function $L(\sigma) = \mathbb{1}[\sigma(i) = 1](\sigma)$ *also* factors riffle independently over the sets A and B. In particular, if $i \in A$, then L factors as the following product of interleaving and relative ranking distributions:

$$L(\sigma) = m_L(\tau_{A,B}(\sigma)) \cdot f_L(\phi_A(\sigma)) \cdot g_L(\phi_B(\sigma_B)),$$

where:

$$m_L(\tau_{A,B}) = \begin{cases} 1 & \text{if } A \text{ occupies the first rank w.r.t. } \tau \\ 0 & \text{otherwise} \end{cases}, \quad f_L(\phi_A) = \mathbb{1}[\sigma(i) = 1](\phi_A), \quad g_L(\phi_B) = 1.$$

As a consequence of the above factorization, the posterior distribution can be obtained by locally conditioning each of the factors in the riffle independent prior distribution and hence can be performed without operating on the entire global distribution. To condition, one iterates through a local factor, zeroing out the interleavings or relative rankings that are not consistent with i mapping to first place, then recursively conditions any descendants that exist. The following result shows that in fact, top-k conditioning in general preserves riffled independence relations and can, in the same way, be performed locally at individual factors.

Theorem 1. *If A and B are riffle independent with respect to a prior distribution h, then A and B are also riffle independent with respect to the posterior distribution $h(\sigma|\sigma^{-1}(1), \sigma^{-1}(2), \ldots, \sigma^{-1}(k))$. More generally, top-$k$ conditioning preserves the structure of a hierarchical riffle independent model.*

Discussion. What is particularly interesting about Thm. 1 is that the same result does *not* hold for observations of arbitrary ranks — for example, it is not easy to condition on a candidate being in 7th place. It is really the structure of riffled independence that uniquely singles out top-k partial rankings as the observations that allow for efficient conditioning by locally modify factors.

4 Structure learning from partial rankings

We now apply our conditioning method to learn the structure and parameters of a hierarchical riffle independent model from top-k ranked data.

Structure learning from full rankings. In the structure learning problem of [6], one finds a hierarchical partitioning of the item set that best corresponds to the riffled independence relationships in a set of rankings drawn i.i.d. from a distribution. In a nutshell, [6] finds subsets of items that are riffle independent by computing an independence measure locally for each triplet of items. These tripletwise independence measures reflect, in a certain sense, the cost of separating some item i from items j and k and can be used to define a clustering-like objective. One shortcoming of the approach, however, is that it relies on a dataset of *full rankings* in order to properly estimate independence.

Expectation-maximization for partially ranked data. To handle partially ranked data, one needs to make additional assumptions about how the partial rankings are generated. In our work, we use a simple *top-k censoring* assumption — that every voter is associated with a latent *full ranking* but reports only the top-k items in his ranking. Alternatively, one can also interpret partial rankings as a vote for all of the permutations that are consistent with a given top-k ranking. We will refer to this second interpretation as the *uniform "fill-in"* assumption and compare against it in our experiments.

Since the rankings beyond the top-k positions can be regarded as latent variables, a simple way to estimate structure and parameters for our riffle independent models is to use the EM algorithm. If hierarchical riffle independent models are analogous to graphical models, then our EM approach is analogous to the *structural EM* algorithms ([2, 3]) that have been developed for graphical model structure learning from incomplete data. In the following, we sketch the two main steps of the algorithm.

E-step: As with typical EM algorithms, the intuition is that if we *knew* the values of the latent variables, then we could estimate parameters and structure. Given a prior distribution h (which can be thought of as an initial setting of structure and parameters), the E-step then "guesses" the fill-in for every top-k ranking $\sigma^{-1}_{1:k} = (\sigma^{-1}(1), \ldots, \sigma^{-1}(k))$ in the training set by computing the posterior distribution $h(\sigma|\sigma^{-1}_{1:k})$ using the procedure given in the previous section.

M-step: In the M-step, one maximizes the expected log-likelihood of the training data with respect to the parameters and structure of the model. While the M-step for structural EM typically involves some sort of local search in the space of graphical structures, the algorithm in [6] is more global in nature and it is not clear if it is possible to modify the algorithm so that it can directly use the expectations computed from the E-step.

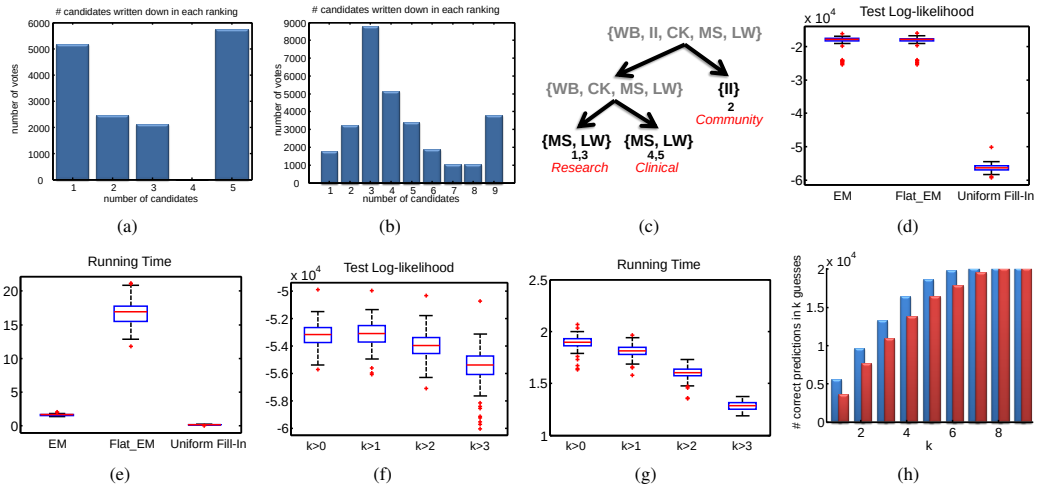

Figure 2: Experiment results

Instead, observing that sampling from riffle independent models can be done exactly and efficiently, we simply sample full rankings from the posterior distributions computed in the E-step and pass these full rankings into the structure learning [6]. The number of samples that are necessary, instead of scaling factorially, scales according to the number of samples required to detect riffled independence (which under mild assumptions is typically polynomial).

5 Experimental evaluation

Datasets. For evaluation, we focus on two datasets. The *APA dataset* [1] is a collection of over 15000 partially ranked ballots from a 1980 presidential election of the American Psychological Association where members rank ordered their top-k of five candidates. Our second dataset, the *Dublin dataset* [4] taken from an Irish House of Parliament election in 2002 with over 24000 rankings of nine candidates. Figures 2(a) and 2(b) plot, for each $k \in \{1, \ldots, 5\}$, the number of people who wrote down k candidates on their ballot for the APA and Dublin datasets respectively. In particular, note that the majority of ballots in both datasets consist of partial rather than full rankings.

Comparison of structure learning strategies. We first compared our EM algorithm against two alternative approaches that we call *FlatEM* and *Uniform Fill-in*. The FlatEM algorithm is the same as the EM algorithm above except for two details: (1) it performs conditioning exhaustively instead of exploiting the factorized model structure, and (2) it performs the M-step without sampling. The Uniform Fill-in approach treats every top-k ranking in the tranining set as a uniform collection of votes for all of the full rankings consistent with that top-k ranking. Figure 2(c) shows the tree that was recovered by the vast majority of 200 bootstrapped runs of our EM method, which happens to be the same tree recovered using only full rankings with leaf nodes corresponding to distinct political coalitions within the APA community. In Figure 2(d) we plot test set loglikelihoods corresponding to each approach, with EM and FlatEM having almost identical results and both performing much better than the Uniform Fill-in approach. On the other hand, Figure 2(e) which plots running times shows that FlatEM can be far more costly (for most datasets, it cannot even be run in a reasonable amount of time).

Evaluating the value of partial rankings. To verify that partial rankings *do* indeed make a difference, we plot the results of estimating a model from the subsets of APA training data consisting of top-k rankings with length larger than some fixed k. Figures 2(f) and 2(g) show the likelihood and running times for $k = 0, 1, 2, 3$ with $k = 0$ being the entire training set and $k = 3$ being the subset of training data consisting only of full rankings. As our results show, including partial rankings does indeed help on average for improving test log-likelihood (with diminishing returns).

Evaluating prediction accuracy. Finally, we present an evaluation on a simple prediction task in which the model is asked to predict the item that a user ranks third given the items in first and second place. We used 5000 rankings from the Dublin dataset for training and plot the number of correct predictions on the remaining data given that the algorithm is allowed k guesses. Figure 2(h) compares the performance of our riffle independent model (blue) against that of a simple first order Markov chain prediction (red) — the baseline method performs surprisingly well, but is still outperformed by our riffle independent model. We conjecture that the difference in performance would be more dramatic if task were to make longer range predictions.

Acknowledgements

We would like to thank Eric Horvitz and Carlos Guestrin for their feedback and discussions, as well as Ryen White and Dan Liebling for help with datasets.

References

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

[2] Nir Friedman. Learning belief networks in the presence of missing values and hidden variables. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 125–133. Morgan Kaufmann, 1997.

[3] Nir Friedman. The bayesian structural em algorithm. In *Uncertainty in Artificial Intelligence (UAI)*, pages 129–138. Morgan Kaufmann, 1998.

[4] Claire Gormley and Brendan Murphy. A latent space model for rank data. In *23rd International Conference on Machine Learning*, 2006.

[5] Jonathan Huang and Carlos Guestrin. Riffled independence for ranked data. In *Advances in Neural Information Processing Systems 23*, 2009.

[6] Jonathan Huang and Carlos Guestrin. Learning hierarchical riffle independent groupings from rankings. In *International Conference on Machine Learning (ICML 2010)*, Haifa, Israel, June 2010.

[7] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[8] G. Lebanon and Y. Mao. Non-parametric modeling of partially ranked data. In *NIPS*, 2008.

[9] M. Steyvers, M.D. Lee, B. Miller, and P. Hemmer. The wisdom of crowds in the recollection of order information. In *Advances in Neural Information Processing Systems, 22*. MIT Press, 2009.