Cluster Analysis of Heterogeneous Rank Data

Ludwig M. Busse BUSSEL@STUDENT.ETHZ.CH
Peter Orbanz PORBANZ@INF.ETHZ.CH
Joachim M. Buhmann JBUHMANN@INF.ETHZ.CH
Institute of Computational Science, ETH Zurich, 8092 Zurich, Switzerland

This revision of the ICML 2007 proceedings article corrects an error in Sec. 3.

Abstract

Cluster analysis of ranking data, which occurs in consumer questionnaires, voting forms or other inquiries of preferences, attempts to identify typical groups of rank choices. Empirically measured rankings are often incomplete, i.e. different numbers of filled rank positions cause heterogeneity in the data. We propose a mixture approach for clustering of heterogeneous rank data. Rankings of different lengths can be described and compared by means of a single probabilistic model. A maximum entropy approach avoids hidden assumptions about missing rank positions. Parameter estimators and an efficient EM algorithm for unsupervised inference are derived for the ranking mixture model. Experiments on both synthetic data and real-world data demonstrate significantly improved parameter estimates on heterogeneous data when the incomplete rankings are included in the inference process.

1. Introduction

Ranking data commonly occurs in preference surveys: A number of subjects are asked to rank a list of items or concepts according to their personal order of preference. Two types of ranking data are usually discussed in the literature: Complete and partial (or incomplete) rankings. A wide range of probabilistic models is available for both (Diaconis, 1988; Critchlow, 1985). A complete ranking of r items is a permutation of these items, listed in order of preference. Mathematical models of rankings are based on the corresponding permutation group. A partial ranking is a preference list of t out of r items. Partial rankings require some refinements of models designed for complete rankings, since two arbitrary partial rankings will in general contain different subsets of the items. An extensive review of rank comparisons can be found in (Critchlow, 1985).

Clustering of rank data aims at the identification of groups of rankers with a common, typical preference behavior (Marden, 1995). An unsupervised clustering method for complete rankings has been proposed in (Murphy & Martin, 2003), based on the well-known Mallows' model (Mallows, 1957) and its generalizations. A different but related problem is the combination of several rankings. This question has recently been discussed by a number of authors, both in Machine Learning (Lebanon & Lafferty, 2002) and discrete algorithmics (Ailon et al., 2005).

For real-world surveys, the data analyst is often confronted with *heterogeneous* data, that is, data containing partial rankings of different lengths. In the well-studied APA data set (Diaconis, 1989), for example, only about a third of the rankings are complete, and the remaining incomplete lists have variable lengths. Common practice in the analysis of heterogeneous rank data is to delete partial rankings, and analyze only the subset of complete rankings (Murphy & Martin, 2003), or to analyze partial rankings of different lengths separately. This raises conceptual problems, as we must expect the removal of a subsample of common characteristic (i.e. incompleteness of the rankings) to cause a systematic bias. Moreover, decreasing the sample size by removing partial rankings can result in a significant decrease of estimation accuracy.

For heterogeneous data, clusters model typical preferences. A ranker associated with any group may either state his preferences completely or incompletely. In other words, each cluster again constitutes a heterogeneous data set, containing rankings of different lengths. As a core contribution of this paper, we

Appearing in *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, OR, 2007.

obtain a model applicable to heterogeneous data by building on the work of (Fligner & Verducci, 1986). The model is a parametric location-scale model based on the Kendall distance (Kendall, 1938), and thus related to Mallows' model (Mallows, 1957). We address the clustering problem by combining several model instances into a parametric mixture. Inference is conducted in a maximum likelihood framework by an expectation-maximization algorithm. The model admits an estimation procedure much more efficient than the straightforward EM approach proposed in the literature for distance-based rank models. Our experiments clearly demonstrate that the additional information in partial rankings can significantly improve parameter estimates of mixture components in rank cluster analysis.

The article is organized as follows: Sec. 2 briefly surveys probabilistic models for complete and partial rank data. A model for heterogeneous rankings is described in Sec. 3, and its algorithmic estimation from data in Sec. 4. Experimental results are presented in Sec. 5.

2. Background

The objective of rank data clustering is to (i) group similar rankings in the input data and (ii) identify rankings that are prototypical representatives for each group. Our approach is probabilistic: A probability model is defined capable of representing an individual group. A mixture of such models is then fitted to the data by an alternating estimation procedure. We will first introduce the standard probability models on rank data available in the literature.

2.1. Models for Complete Rankings

We assume that rank data for r items are observed. The items are indexed $m = 1, \ldots, r$, and n subjects are asked to arrange the items according to their order of preference. Each of the resulting lists can be regarded as a permutation π_i of the item indices, i.e. $\pi_i(m) = j$ indicates that the i-th ranker has assigned rank j to item m. The set of possible rankings is then given by the set of possible permutations of r items. This set has a group structure and is referred to as the *symmetric group* of order r, denoted $\mathbb{S}(r)$.

Statistics has developed a sizable amount of rank data models. Of particular interest for data clustering are the so-called *distance-based models* of the form

$$P(\pi|\lambda, \sigma) := \frac{1}{Z(\lambda)} \exp\left(-\lambda d(\pi, \sigma)\right) \ , \qquad (1)$$

with $Z(\lambda) := \sum_{\pi \in \mathbb{S}(r)} \exp\left(-\lambda d(\pi, \sigma)\right)$. The model is parameterized by a ranking $\sigma \in \mathbb{S}(r)$ and a dispersion parameter $\lambda \in \mathbb{R}_+$. The function $d : \mathbb{S}(r) \times \mathbb{S}(r) \to \mathbb{R}_{\geq 0}$ is a *distance function*, i.e. a function with metric properties on $\mathbb{S}(r)$. Since d is a metric and hence $d(\pi, \sigma) = 0$ iff $\pi = \sigma$, the distribution P assumes its unique mode at σ, and σ is referred to as the *modal ranking*. The dispersion parameter λ controls how sharply the distribution peaks around the mode, i.e. small (large) λ values code for broad (peaked) distributions. For clustering, distance-based models capture the notion that two observations belong to the same group if they are "close". The approach is related to familiar clustering methods for other data types, such as Gaussian mixtures for vectorial data (which measure distance by Euclidean or covariance-adjusted Euclidean distance) or multinomial mixtures for histogram data (which measure a distance-like quantity by Kullback-Leibler divergence). Different models can be obtained by substituting different types of metrics for d in (1). Other popular choices include the Spearman rank correlation metric, and the Hamming, Cayley and Ulam distances (Critchlow, 1985). The present work focuses on one metric in particular, the widely used *Kendall distance* (Kendall, 1938), defined as

$$d_\tau(\pi, \sigma) \quad := \quad \text{minimum of adjacent transpositions required to transform } \pi \text{ into } \sigma.$$

Closely related is the *Cayley distance*, which drops the adjacency requirement, and thus measures the distance in terms of arbitrary transpositions. For $d = d_\tau$, the model (1) is *Mallows' ϕ model* (Mallows, 1957) in its original form. More generally, models of the form (1) are usually referred to as Mallows models, provided that d is a metric.

2.2. Clustering with Mallows' Model

For clustering, the observed rank data is assumed to consist of K groups. Each group is modeled by a Mallows distribution

$$P_k(\pi|\lambda_k, \sigma_k) := \frac{1}{Z(\lambda_k)} \exp\left(-\lambda_k d_\tau(\pi, \sigma_k)\right) \ . \qquad (2)$$

The component distributions are joined in a mixture model,

$$Q(\pi) := \sum_{k=1}^{K} c_k P_k(\pi|\lambda_k, \sigma_k) \ , \qquad (3)$$

where the mixture weights $(c_1, \ldots, c_K)$ form a partition of 1. Model parameters can be estimated with an expectation-maximization (EM) algorithm (McLachlan & Krishnan, 1997), or more sophisticated latent variable estimation algorithms, such as Simulated Annealing or Deterministic Annealing (Kirkpatrick et al., 1983; Hofmann & Buhmann, 1997).

2.3. Partial Rankings

A *partial ranking* is a ranking of t out of r items. Usually, one assumes a top-t ranking, i.e. subjects have ranked their t favorites out of a larger number of r items. Distance-based models for partial rankings can be constructed by generalizing metrics on complete rankings to valid metrics on partial rankings. (Critchlow, 1985) has proposed such a generalization based on Hausdorff distances.

A partial top-t ranking is best represented as an inverse: In standard notation, regarding the permutation π as a list of numbers, position in the list corresponds to an item index (and the entry value at that position gives a rank). A ranking of t favorite items is thus a list with gaps. Written as the inverse π^{-1}, position denotes rank, and a top-t ranking has the form $\pi^{-1} = (\pi^{-1}(1), \ldots, \pi^{-1}(t), *, \ldots, *)$. For any partial ranking π of length t, denote by $C(\pi)$ the set of all complete rankings $\tilde{\pi}$ matching π in their first t positions, that is, $C(\pi) := \{\tilde{\pi} \in \mathbb{S}(r) | \tilde{\pi}(j) = \pi(j), j = 1, \ldots, t\}$. We will refer to C as the *consistent set* of π (in algebraic terms, this is just the right coset $\mathbb{S}_{r-t}\pi$). For any two different partial rankings of the same length, the consistent sets are disjoint, and their union over all partial rankings of a given length is $\mathbb{S}(r)$. For a given metric d on $\mathbb{S}(r)$, (Critchlow, 1985) defines an induced metric d^* on partial rankings as the Hausdorff distance between their consistent sets. As put by Critchlow, $d^*(\pi, \sigma)$ can be imagined as the smallest amount by which $C(\pi)$ has to be enlarged to include all of $C(\sigma)$. Another approach to partial rankings is the completion method proposed by (Beckett, 1993), who estimates complete rankings from partial ones based on a Mallows model (cf. Sec. 5).

3. Modeling Heterogeneous Data

In the present work, we consider the problem of modeling real-world survey data, which usually includes partial rankings of variable length t. Differences arise because many subjects will rank only their favorite t items. For ranking data on r items, we therefore have to assume an observed sample to contain partial rankings of all possible lengths $t = 1, \ldots, (r-1)$ (note that $t = (r-1)$ is equivalent to $t = r$, since the missing position is uniquely determined). [1]

[1] We do not consider partial rankings with gaps, i.e. rankings with a total of $t < r$ filled position and empty ranks in between, since data of this type can be expected to be rare. Our model does, in principle, generalize to the case of rankings with gaps, but the actual computations become more difficult.

3.1. Choice of Metric

The model described in this section is based on the Kendall distance. Our choice of the metric is motivated by a range of properties: First, it has an intuitive and plausible interpretation as a number of pairwise choices. (Mallows, 1957) argues that it provides the best possible description of the process of ranking items as performed by a human. Second, it enjoys a high de-facto relevance due to its widespread use. Third, there is a number of appealing mathematical properties: It counts (rather than measures), is efficiently computable, decomposable into a sum, and its standardized distribution has a normal limit (Diaconis, 1988). Though our study is limited to the Kendall case, Fligner-Verducci type models can be derived for the Cayley distance as well (Fligner & Verducci, 1986).

3.2. Probabilistic Model

If only a subset of the available items is ranked, the choice of a probabilistic model implies a distribution assumption for the missing entries. We take a maximum entropy approach, demanding our model to be maximally noncommittal with respect to the missing information. Such a model is suitable to address several generative scenarios for partial rankings: One is indifference of the ranker, i.e. a subject ranks t favorite items, but does not have any preferences concerning the remainder. Another setting are large sets of items, where most subjects will not take the time to provide a complete list (e.g. when the task is to specify a ranking of favorites out of thousands of items). In general, the approach is applicable unless prior information on the popularity of items is available. A maximum entropy approach is optimal in the sense that it does not introduce implicit (hidden) assumptions on the choice of items. This is a notable difference to the Hausdorff metric approach, for example, which constitutes a worst-case assumption: The distance problem is reduced to the original metric by expanding a pair of partial rankings into that consistent pair of complete rankings which differs most under the inducing metric.

To express lack of knowledge w.r.t. to items beyond the preferred t choices, we have to assume that the ranker's choice effectively encompasses all possible completions of π to a complete ranking in $\mathbb{S}(r)$. In other words, successive ranking of items is regarded as a constraining process: By each additional item entered into the list, the ranker constrains the set of possible completions. A full ranking limits $\mathbb{S}(r)$ down to a single element. A partial ranking defines the set $C(\pi)$ of possible completions. Any model distribution P on complete rankings can then be generalized to a distribution P^t on par-

tial rankings by defining the probability of π under P^t as the total probability placed on the set $C(\pi)$ by the model P:

$$P^t(\pi) := P(C(\pi)) = \sum_{\tilde{\pi} \in C(\pi)} P(\tilde{\pi}) \ . \tag{4}$$

For Mallows' model based on the Kendall distance, the probability $P(C(\pi))$ admits an elegant decomposition. From a statistics point of view, the approach can be regarded as a censored data problem. For the Kendall metric, censored rank data has been considered in (Fligner & Verducci, 1986). They build on the well-known fact that the Kendall distance, as well as the Cayley and Hamming distances, can be decomposed into a sum. Define the following statistic for each position $j = 1, \ldots, (r-1)$ in a complete ranking π of length r:

$$\tilde{s}_j(\pi) := \sum_{l=j+1}^{r} I\{\pi^{-1}(j) > \pi^{-1}(l)\} \ , \tag{5}$$

where π^{-1} denotes the inverse of π in $\mathbb{S}(r)$ and I the indicator function of a set. Intuitively, $\tilde{s}_j$ is the number of adjacent transpositions required to move item j to position j, if the items at the previous $1, \ldots, (j-1)$ are already ordered. The sum over the statistics $\tilde{s}_j$ is the Kendall distance of π and the identity permutation $\mathrm{Id}_{\mathbb{S}(r)}$ (Fligner & Verducci, 1986). The metric d_τ is *right-invariant*, that is, for any $\pi_1, \pi_2, \pi_3 \in \mathbb{S}(r)$, $d_\tau(\pi_1\pi_3, \pi_2\pi_3) = d_\tau(\pi_1, \pi_2)$. Hence, for any $\sigma \in \mathbb{S}(r)$,

$$d_\tau(\pi, \sigma) = d_\tau(\pi\sigma^{-1}, \mathrm{Id}_{\mathbb{S}(r)}) = \sum_{j=1}^{r-1} \tilde{s}_j(\pi\sigma^{-1}) \ . \tag{6}$$

This representation is somewhat inconvenient for modeling partial rankings, since the sum ranges over the suffix of rank j, which includes empty positions. We therefore substitute equivalent statistics s_j involving only indices up to j. For any permutation ρ, define

$$s_j(\rho) := \rho(j) - \sum_{l=1}^{j} I\{\rho(j) \geq \rho(l)\} \ . \tag{7}$$

The Kendall metric is then computed as $d_\tau(\pi, \sigma) := \sum_{j=1}^{r} s_j(\sigma\pi^{-1})$, which avoids any explicit use of π: Since π^{-1} is a top-t list, it is not invertible. The importance of the sum representation for modeling partial rankings is that it can be decomposed into terms corresponding to filled and empty positions, respectively:

$$\begin{aligned} d_\tau(\pi, \sigma) &= \sum_{j=1}^{t} s_j(\sigma\pi^{-1}) + \sum_{j=t+1}^{r} s_j(\sigma\pi^{-1}) \\ &= \mathbf{s}^t(\sigma\pi^{-1}) + \mathbf{s}^{\mathrm{empty}}(\sigma\pi^{-1}) \ . \end{aligned} \tag{8}$$

The probability of the consistent set of π under Mallows' model can then be expressed as

$$\begin{aligned} P(C(\pi)|\lambda, \sigma) &= \frac{1}{Z(\lambda)} \sum_{\tilde{\pi} \in C(\pi)} \exp\left(-\lambda d_\tau(\tilde{\pi}, \sigma)\right) \\ &= \frac{\exp\left(-\lambda \mathbf{s}^t(\sigma\pi^{-1})\right)}{Z(\lambda)} \sum_{\tilde{\pi} \in C(\pi)} \exp\left(-\lambda \mathbf{s}^{\mathrm{empty}}(\sigma\tilde{\pi}^{-1})\right) \end{aligned}$$

The sum over $C(\pi)$ depends only on t, and is absorbed into the partition function $Z(\lambda)$. Hence, the resulting partition function $Z^t(\lambda)$ depends on t. The probability of the partial ranking is thus

$$P(C(\pi)|\lambda, \sigma) = \frac{1}{Z^t(\lambda)} \exp\left(-\lambda \mathbf{s}^t(\sigma\pi^{-1})\right) \ , \tag{9}$$

and we write $P(\pi|\lambda, \sigma) := P(C(\pi)|\lambda, \sigma)$. The partition function Z^t can be derived from the (somewhat more complicated) model in (Fligner & Verducci, 1986), as

$$Z^t(\lambda) := \prod_{j=1}^{t} \frac{1 - e^{-\lambda(r-j+1)}}{1 - e^{-\lambda}} \ . \tag{10}$$

The distribution is a maximum entropy model, as it constitutes an exponential family distribution given the modal ranking σ, with the functions s_j as its sufficient statistics. The choice of the location parameter σ does not change the model's entropy.

Heterogeneous, partial ranking data drawn from K distinct groups can now be described by a mixture model. Denote by $t(\pi)$ the length of an arbitrary partial ranking π. The generative model for the data is then

$$Q(\pi|\mathbf{c}, \boldsymbol{\lambda}, \boldsymbol{\sigma}) := \sum_{k=1}^{K} \frac{c_k}{Z^{t(\pi)}(\lambda_k)} e^{-\lambda_k \mathbf{s}^{t(\pi)}(\sigma\pi_k^{-1})} \ . \tag{11}$$

To summarize, lack of knowledge (or indifference of a ranker) about the order of neglected items is expressed by substituting the consistent set of a ranking in the modeling process. Probabilities are comparable for rankings of different lengths. Formally, this holds because the model is a distribution on the consistent sets $C(\pi)$. For any two rankings, the sets are nested if one ranking prefixes the other, and are disjoint otherwise. The mixture expresses the separation of the rankers surveyed in the data into different groups or types, each of which exhibits a "typical" preference behavior. The data collected from rankers within a single group will in general be heterogeneous. For a given group, the modal ranking describes a consensus preference, and the corresponding dispersion parameter variation between the associated rankers.

4. Model Inference

Our approach to inference is based on maximum likelihood (ML) estimation. For the mixture model described above, the overall ML estimator of the model parameters is approximated with an expectation-maximization (EM) algorithm (McLachlan & Krishnan, 1997). In this section, we derive estimation equations for the heterogeneous data model, and discuss the implementation of efficient EM algorithms for rank data. Straightforward implementations of such algorithms previously proposed for Mallows mixtures on complete rankings (Murphy & Martin, 2003) require the repeated evaluation of sums over all possible rankings. Since the symmetric group $\mathbb{S}(r)$ has $r!$ elements, such methods are only applicable for rankings with a small number of entries.

For data $\pi_i, i = 1, \ldots, n$ and K clusters, we define binary class assignment vectors $\mathbf{M}_i := (M_{i1}, \ldots, M_{iK})$. If π_i is assigned to cluster k, then $M_{ik} = 1$ and all other entries are set to zero. These are the hidden variables of the EM estimation problem. The EM algorithm relaxes the binary assignments to assignment probabilities $q_{ik} := \mathrm{E}[M_{ik}]$, where $q_{ik} \in [0, 1]$ and $\sum_k q_{ik} = 1$ for each i. The E-step of the algorithm computes estimates of the assignment probabilities conditional on the current parameter configuration of the model. Given estimates of the component parameters λ_k, σ_k and the mixture weight c_k for each cluster k, assignment probabilities are estimated as $q_{ik} := \frac{c_k P^t(\pi_i|\lambda_k,\sigma_k)}{\sum_{l=1}^K c_l P^t(\pi_i|\lambda_l,\sigma_l)}$. In the M-step, assignment probabilities are assumed to be given. For each cluster, the parameters to be estimated are c_k, λ_k and σ_k. As for any mixture model EM algorithm, the mixture weights are straightforwardly computed as $c_k := \frac{1}{n}\sum_{i=1}^n q_{ik}$. ML estimation of the component parameters σ_k, λ_k proceeds in two steps, first obtaining an estimate of σ_k (which does not depend on λ_k), and then estimating λ_k conditional on σ_k. This is reminiscent of e.g. the two-stage ML estimation of location and scale parameters for Gaussian models. The modal ranking ML estimate is

$$\begin{aligned}\hat{\sigma}_k &= \arg\max_{\sigma_k} \log \prod_{i=1}^n P(\pi_i^t|\lambda_k, \sigma_k)^{q_{ik}} \\ &= \arg\min_{\sigma_k} \sum_{i=1}^n q_{ik} \sum_{j=1}^{t(\pi_i)} s_j(\sigma_k \pi_i^{-1}) \,. \qquad (12)\end{aligned}$$

Rather than evaluating the minimum over the whole group, our algorithm performs a local search step, by minimizing over all adjacent transpositions around the estimate $\tilde{\sigma}_k$ obtained during the previous M-step. This strategy is equivalent to searching within a d_τ-radius of 1. When initialized at random, the algorithm may thus require several steps until it reaches the correct σ_k. The local search results in a generalized EM (GEM) algorithm, since the conditional likelihood is increased but not fully maximized during the M-step. Generalized EM algorithms satisfy the EM convergence conditions and retain EM convergence guarantees (McLachlan & Krishnan, 1997). Our control experiments in Sec. 5 clearly indicate that the local estimation approach is adequate. If modal ranking estimation errors occur, they are due to ambiguous data, i.e. data drawn from clusters for which the distance between the modal rankings is small w.r.t. to their dispersion. Local search over transpositions reduces the estimation costs for σ_k from $r!$ to r evaluations.

Since the dispersion parameter is continuous, a maximum condition for the likelihood w.r.t. λ_k can be obtained by differentiation. Setting the derivative of the log-likelihood of one mode to zero yields

$$-\sum_{i=1}^n \frac{\partial}{\partial \lambda_k} \log Z^{t(\pi_i)}(\lambda_k) = \sum_{i=1}^n d(\pi_i, \sigma_k) \,. \qquad (13)$$

For our heterogeneous data model as described in Sec. 3, (i) the partition function has a closed-form solution and the derivative can be obtained explicitly, and (ii) the model has to be decomposed over different types of rankings, since the partition function depends on t. Assume that the observations π_i have different lengths $t \in \{1, \ldots, r\}$. Denote by $I_t \subset \{1, \ldots, n\}$ the set of indices i for which π_i has length t. The log-likelihood of the complete data set under cluster k is

$$\begin{aligned}\log \prod_{i=1}^n P(\pi_i|\lambda_k, \sigma_k) &= \sum_{t=1}^r \sum_{i \in I_t} \log P(\pi_i|\lambda_k, \sigma_k) \\ = -\sum_{t=1}^r |I_t| \log(Z^t(\lambda_k)) &- \sum_{t=1}^r \sum_{i \in I_t} \lambda_k \sum_{j=1}^t s_j(\sigma_k \pi_i^{-1})\end{aligned}$$

Equating the derivative to zero gives

$$-\sum_{t=1}^r |I_t| \frac{\partial}{\partial \lambda_k} \log(Z^t(\lambda_k)) = \sum_{i=1}^n \sum_{j=1}^{t(\pi_i)} s_j(\sigma_k \pi_i^{-1}) \,. \qquad (14)$$

The derivative of $\log(Z^t(\lambda_k))$ for given t is

$$\frac{\partial}{\partial \lambda_k} \log(Z^t(\lambda_k)) = \sum_{j=r-t+1}^r \frac{j}{e^{j\lambda_k} - 1} - \frac{t}{e^{\lambda_k} - 1} \,.$$

This expression is both rapidly computable and smooth w.r.t. λ_k. The right hand side of (14) does not depend on λ_k, hence the maximum likelihood estimator $\hat{\lambda}_k$ can be efficiently evaluated by numerical solution of equation (14).

Table 1. Estimation errors on artificial data of sample size $n = 300$, with $K = 3$ clusters. For uniform c, all clusters have equal size. For non-uniform c, cluster sizes differ.

Settings			Results		
c	d	λ	$\hat{K}$	error $\hat{c}$	error $\hat{\lambda}$
uniform	[2, 9, 9]	0.50	1	0.033	0.086
		1.00	3	0.007	0.056
		1.50	3	0.027	0.151
	[8, 6, 6]	0.50	1	0.155	0.274
		1.00	3	0.029	0.094
		1.50	3	0.016	0.050
non-uniform	[2, 9, 9]	0.50	1	0.248	0.324
		1.00	3	0.013	0.032
		1.50	3	0.001	0.048
	[8, 6, 6]	0.50	1	0.189	0.331
		1.00	3	0.047	0.144
		1.50	3	0.013	0.057

5. Experimental Results

The experiments include artificial and real-world rank data. The mixture analysis with artificial data drawn from a density with known parameters is conducted to evaluate the algorithm's effectiveness in recovering parameters from rank data. Additional experiments are conducted on the American Psychological Association (APA) data set (Diaconis, 1989). All experiments are performed with the EM algorithm described in Sec. 4. The number of clusters is selected by a Bayesian Information Criterion (BIC) (McLachlan & Krishnan, 1997). For comparison, we use a clustering approach based on the completion method described in (Beckett, 1993). The method explicitly estimates a maximum likelihood completion to a full ranking by treating the missing positions as latent information, and assuming complete rankings to be distributed according to a Mallows model. An estimate of the full ranking is obtained with an EM algorithm, which alternatingly estimates a Mallows model from current completion estimates, and then estimates completions based on the current model. The method can be used as basis for partial rank data clustering model, by performing completions based on the data currently assigned to a cluster during the clustering E-step, and performing maximum likelihood estimation for the mixture components given the current completion estimates during the M-step.

Table 2. Long rankings: Estimation error comparison for ranking length $r = 20$, with $K = 10$ clusters and $n = 1000$ samples (uniform over partial lengths).

Method	error $\hat{\sigma}_k$	error $\hat{\lambda}_k$
Maximum Entropy	0	0.06 ± 0.01
Beckett's completion	1.52 ± 0.57	0.11 ± 0.02

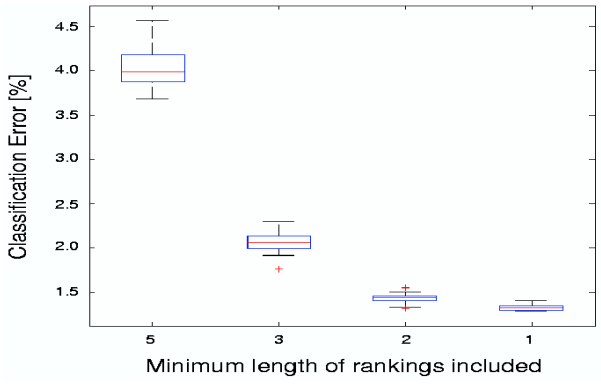

Figure 1. Full versus restricted data set: Average estimation error for cluster assignments (vertical) versus the number of ranking types present in the data set (horizontal).

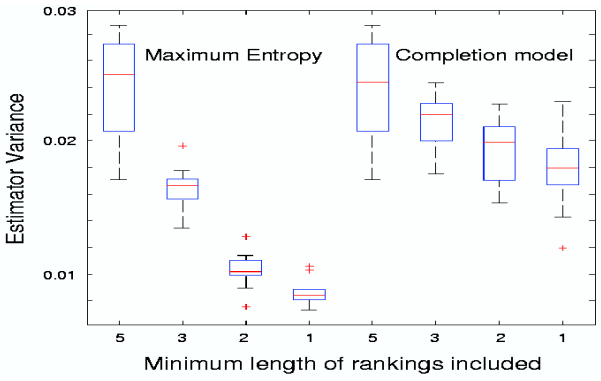

Figure 2. APA data set: Variance of dispersion estimates (vertical) versus number of ranking types present in the data set (horizontal), for our method (left) and Beckett's completion model (right). Minimum length 5 corresponds to the subset of complete rankings, 1 to the whole data set. The variance is computed over 20 bootstrap samples.

5.1. Synthetic Data

Synthetic data observations were drawn at random from a mixture model (11). Sample experiments for $r = 5$ items and $K = 3$ clusters are shown in Tab. 1. By d, we denote the pairwise Kendall distances between the cluster centers. The quality of parameter estimates is reported as mean squared error on $n = 300$ observations. The BIC estimate $\hat{K}$ of the number of clusters is accurate except for very small λ which corresponds to broad modes. This behavior is expected since the different modes strongly overlap for

small λ and, consequently, are not resolvable for the chosen number of observations. When BIC underestimates the number of clusters, the estimation errors for λ and c generally increase. Estimation errors increase again for $\lambda = 1.5$ in the case of two close clusters ($d = [2, 9, 9]$), a distortion effect caused by points of the neighboring cluster. The dispersion at which the effect becomes visible depends on a trade-off between the dispersion and the distance of the clusters. It will occur at a larger value of λ if the clusters are closer. Remarkably, the modal rankings σ_k are always estimated correctly, unless the estimate of the cluster number is wrong.

The value of partial rankings for estimation is illustrated by Fig. 1. EM estimation of the mixture model was conducted on a random data set, with $r = 5$ and a proportion of 25% complete rankings. The partial rankings of lengths $\{1, 2, 3\}$ are also drawn with probability 0.25 each. The estimation error for the cluster assignments was recorded and plotted against the number of ranking length types present in the data (horizontal), where 5 denotes the case where all partial rankings are removed from the data set, corresponding to the common practice of analyzing only the subset of complete rankings. When more categories are added (with 1 corresponding to the complete heterogeneous data set), we observe a significant decrease in both the estimation error and its variance. A double-logarithmic plot of these results reveals an approximate scaling behavior of $\mathcal{O}(1/\sqrt{n})$. We conclude that, at least in the controlled setting of synthetically generated data, the inference procedure is capable of using the information carried by partial rankings to its advantage.

Comparisons with Beckett's completion method were conducted for rankings of length $r = 5$ and $r = 20$ on synthetic data. Parameter estimates obtained by our method are more accurate then those obtained by the completion approach. The difference is statistically significant even for $r = 5$, and becomes more pronounced as the number of items is increased. Results for $r = 20$ are reported in Tab. 2. Application of Beckett's method to rankings of this length requires a modification of the original algorithm. Beckett's estimation step completely enumerates the consistent set of each partial ranking, and hence scales exponentially in the number of unranked items. It can be made applicable to large rankings by substituting a sampling step, at the price of an increase in the variance of estimates. The completion method introduces an error in the estimation of the modal ranking. Errors are caused by the large number of latent variables required by the completion model, which result in diffuse distributions of the cluster assignments.

5.2. APA Data

The APA data set of real-world rankings was obtained from the results of the American Psychological Association's 1980 presidential election. Each ballot is a ranking of five candidates. The data set is remarkably large (about 15,000 observations) and it has been extensively analyzed (Diaconis, 1988). The data is heterogeneous, that is, only 5738 ballots contain complete rankings. The remainder contains top-t rankings of all possible lengths $t = 1$ through $t = 3$ (note that $t = 4$ is equivalent to a complete ranking). Since no ground-truth is available for this data, the estimation errors cannot be computed. However, to analyze the value of the partial rankings for estimation accuracy, we consider the variance of the estimate of λ. Fig. 2 shows a plot of the bootstrap variance estimate of the estimators $\lambda_1, \ldots, \lambda_K$, for both our model and clustering based on Beckett's completion approach. The variance estimates are plotted versus the number of ranking types (i.e. different lengths). The error bars meausure variances over multiple repetitions of the bootstrap estimation experiment. For our maximum entropy model (left), inclusion of additional partial observations in the analysis clearly stabilizes parameter estimates. The variance remains notably higher for the Beckett approach (right). Using Beckett's completion requires latent variables to account for the missing positions, in addition to the assignment variables required by the mixture model. Since additional latent variables increase the overall entropy of the model, the completion approach has a destabilizing effect, which becomes more pronounced as the proportion of partial rankings in the data increases. It will also slow down convergence of the inference algorithm, as the convergence speed of EM algorithms depends on the proportion of latent variables (McLachlan & Krishnan, 1997).

6. Conclusion

We have presented an unsupervised clustering approach for ranking data that is capable of performing an integrated analysis on heterogeneous, real-world data, rather then decimating the data to fit the model. An efficient EM algorithm has been derived and shown to recover parameters accurately from data.

Our method offers two advantages compared to rank data clustering techniques available in the literature: (i) the ability to analyze a data set composed of different ranking types, and (ii) efficient inference. The value of the former point was demonstrated by our

experiments: Removing partial rankings from a given data set significantly reduces the accuracy of parameter estimates. For data containing only complete rankings, a decrease in estimation accuracy would have to be expected if samples are removed. That the same effect is observable (Fig. 1) when the removed rankings are partial shows that incomplete rankings carry valuable information – even those containing only a single entry.

However, on real-world survey data, this effective loss in sample size is not the only consequence of removing data. In a survey, ranking only partially may constitute a typical behavior. That is, if providing a partial rather than a complete ranking correlates with certain preferences, removing partial rankings will exclude these modes of behavior from the analysis. In addition to reducing the sample size, it also introduces a systematic bias. Both drawbacks can be avoided by automatic analysis methods capable of processing heterogeneous data, and combining estimate contributions obtained on rankings of different lengths in a meaningful way. Our modeling approach permits the natural integration of different length types by defining a distribution on the subset of completions consistent with a given partial ranking.

Algorithmic inference of our model is substantially more efficient than the algorithms available in the literature for distance-based models. The EM algorithm presented in Sec. 4 scales linearly in the number of ranked items (i.e. the order r of the permutation group), rather than exponentially, as other algorithms do (Murphy & Martin, 2003).

Our modeling approach relies on the decomposition of the Kendall distance into a sum over ranking positions and, therefore, it generalizes to ranking metrics with the same property. Such a decomposition is known for the Kendall, Cayley and Hamming distances, but results from Weyl group theory suggest that it does not exist for other metrics (Diaconis, 1988). Approximate decompositions for other metrics, however, might render efficient relaxations possible which would generalize our approach to these cases. Our emphasis on the Kendall metric is motivated by its ubiquitous usage in rank mixture analysis and by its natural properties (see Sec. 3.1) for rank comparisons.

Acknowledgments

We thank Volker Roth and the anonymous reviewers for helpful suggestions, and Marina Meila for pointing out an error in Sec. 3.

References

Ailon, N., Charikar, M., & Newman, A. (2005). Aggregating inconsistent information: Ranking and clustering. *ACM Symposium on the Theory of Computing*.

Beckett, L. A. (1993). Maximum likelihood estimation in Mallows' model using partially ranked data. In M. A. Fligner and J. S. Verducci (Eds.), *Probability models and statistical analyses for ranking data*.

Critchlow, D. (1985). *Metric methods for analyzing partially ranked data*. Springer.

Diaconis, P. (1988). *Group representations in probability and statistics*. Institute of Mathematical Statistics.

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

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

Hofmann, T., & Buhmann, J. (1997). Pairwise data clustering by deterministic annealing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*, 1–14.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*, 81–93.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*, 671–680.

Lebanon, G., & Lafferty, J. (2002). Cranking: Combining rankings using conditional probability models on permutations. *International Conference on Machine Learning*.

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

Marden, J. I. (1995). *Analyzing and modeling rank data*. Chapman & Hall.

McLachlan, G. J., & Krishnan, T. (1997). *The EM algorithm and extensions*. John Wiley & Sons.

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