Model-Based Clustering of Inhomogeneous Paired Comparison Data

Ludwig M. Busse and Joachim M. Buhmann

Department of Computer Science, ETH Zurich,
8092 Zurich, Switzerland
{ludwig.busse,jbuhmann}@inf.ethz.ch

Abstract. This paper demonstrates the derivation of a clustering model for paired comparison data. Similarities for non-Euclidean, ordinal data are handled in the model such that it is capable of performing an integrated analysis on real-world data with different patterns of missings.

Rank-based pairwise comparison matrices with missing entries can be described and compared by means of a probabilistic mixture model defined on the symmetric group. Our EM-method offers two advantages compared to models for pairwise comparison rank data available in the literature: (i) it identifies groups in the pairwise choices based on similarity (ii) it provides the ability to analyze a data set of heterogeneous character w.r.t. to the structural properties of individual data samples.

Furthermore, we devise an active learning strategy for selecting paired comparisons that are highly informative to extract the underlying ranking of the objects. The model can be employed to predict pairwise choice probabilities for individuals and, therefore, it can be used for preference modeling.

1 Introduction

Objects $o_a, o_b, \ldots$ of a given set of objects $\mathcal{O}$ can be characterized in the most elementary form by a preference relation. Such pairwise comparisons, that yield so-called paired comparison data, encode the preferences of objects in many different contexts. Comparing two objects o_a and o_b with an operator $<$, i.e. measuring whether object o_a is bigger, higher, more preferred, ... than object o_b endows an otherwise unstructured pair of objects with a very elementary piece of information (1-bit). Neither the actual difference between the two objects is important nor are there any compulsary restrictions placed on the operator (e.g. transitivity). The data type is a comparison matrix, where objects $o_a, o_b, \ldots, o_m$ are compared pairwise to each other:

$$\mathbf{X} = (x_{ab}) \in \mathbb{B}^{m \times m} = \{0, 1\}^{m \times m}$$

The comparison operator can be specified dependent on the application at hand. Here, we focus in particular on preference data.

A data set consists of $i = 1, \ldots, n$ samples:

$$\mathcal{X} = \left\{ \mathbf{X}^{(i)} \right\} = \left\{ (x_{ab})^{(i)} \right\}$$

M. Pelillo and E.R. Hancock (Eds.): SIMBAD 2011, LNCS 7005, pp. 207–221, 2011.

In this work, we aim at finding structure in pairwise comparison rank data. A famous simple model for pairwise comparisons is the Babington Smith model (a thorough overview is provided in [19]).

Many data arise as pairwise comparisons (rather than as points in an Euclidean vector space $\mathbb{R}^n$). Suppose we have a number of objects o_a, o_b, ... which are to be considered according to some common quality. If the quality is measurable in some objective way, the objects yield variate values, and the problem is amenable to treatment by standard machine learning methods. However, it may happen for either theoretical or for practical reasons that the quality is not measurable or cannot be measured robustly. We then have to rely for a discussion of the variation of the quality based on a comparison of the objects among themselves. The method of pairwise comparisons provides reliable and informative data about the relative quality[1].

A widely used methods of comparison ranks the object according to a suitable application criterion. The objects are arranged in the order in which they possess the quality under consideration (*total order*). The ranking method is not appropriate [14] when the quality considered is not known to be representable by a linear variable. It is not necessarily unreasonable that object $o_a < o_b$, $o_b < o_c$ and $o_c < o_a$, if the objects deal for example with tastes in music, eatables or film stars; and in practice this in not uncommon [14]. Such "inconsistent" information can never appear in a ranking for if o_a is preferred to o_b and o_b to o_c, then o_a must automatically be shown as preferred to o_c. The use of rankings thus destroys what may be valuable information.

When preference relations are evaluated under a single criterion, there is one dominant total order (ranking assumption). However, noise can result in probabilistically intransitive data. In this paper, we consider a probabilistic model for pairwise comparison data, establishing a probability distribution over rankings. The model allows for intransitivities and places equal probability mass on all rankings that are equally consistent with the given pairwise comparisons. Noisy real-world data can be handled in a meaningful way.

[1] Data derived from paired comparison experiments: Many situations naturally produce pairwise comparisons such as sporting events which involve two teams (e.g. football, basketball). The records of wins and losses for the teams constitute the data. In other situations, such as food tasting, pairwise comparisons are helpful because of the difficulty of distinguishing preferences when more than two objects are considered simultaneously. Though direct *rankings* are popular to elicit *preferences* e.g. in music, movies, and food, giving a ranking for more than, say, 5 objects is quite a difficult and time-consuming task for an interviewee to complete. Deciding between just 2 objects at a time is easier than infering complete rankings and thus, pairwise comparison generates data of superior quality. An alternative to asking someone to rank the m objects is to have the ranker choose which of each pair of objects is preferred. With many objects being up for consideration (e.g. products), we must expect the stated pairwise preference data to have missings. Pairwise comparison matrices might be incomplete because respondents do not express all preferences or are indifferent.

We derive a mixture model for cluster analysis and provide an EM algorithm for parameter estimation (unsupervised inference). The model and parameter estimator can tolerate missings in the data, incase not all paired comparisons are made or available to the data analyst.

We also devise a strategy for automatically selecting paired comparisons that are "significant" to extract a ranking.

The model framework introduced above is instantiated for the application of *preference modeling*. Cluster analysis of paired comparison data attempts to find groups of preference choices. Preference data of surveys often suffer from missing values since respondents might answer to only a few paired comparisons, possibly a different set of paired comparisons for each respondent causing heterogeneity in the data. We present a mixture approach for similarity-based pattern analysis of such discrete, non-Euclidean, and inhomogeneous preference data by a single probabilistic model. The usefulness of the method is demonstrated by that predictions (=recommendations) for individuals can be made based on the cluster solution.

This paper is organized as follows: A model for heterogeneous paired comparison data comprising different clusters and missings is presented in Sec. 3, and its algorithmic estimation from data in Sec. 4. Sec. 5 proposes a strategy for selecting pairwise comparisons. In Sec. 6, we point out that the method is helpful for predicting preferences. Experimental results are reported in Sec. 7.

2 Relevant Work

Learning to rank and *ordinal regression* are presently popular research topics. In [7], the problem of learning how to order instances, given feedback in the form of preference judgments, is tackled. Another *supervised* approach to learning a preference function is [10]. Here, the training information consists of samples with partial and possibly inconsistent information about their associated rankings. From these, a ranking function is induced. Learning a preference function, defined over pairs, for producing a ranking is also presented in [2]. An approach to ensemble learning is introduced in [15], which takes ranking rather than classification as fundamental. Multiple input rankings are combined according to the degree of expertise that each ranker has. A supervised pairwise/listwise approach to ranking is developed in [6], and in [21], the problem of consensus finding for a group of rankers is considered.

Unsupervised learning on rank and pairwise data is mostly considered in the context of Collaborative Filtering (see [23] for a survey of techniques). A model for the cluster analysis of rank-type data is developed in [4], which is now relaxed to accomodate for paired comparison data. Learning Mallows models with pairwise preferences was very recently developed in [17].

3 Modeling Paired Comparison Data

When modeling paired comparison data there are two options: i) To model the pairwise comparison process (physical/mechanical/behavioral/neurological processes). ii) To model the population of n pairwise comparison givers (comparators). Here, we focus in this second, data-analytic approach.

Suppose there are m *objects*, also called *items*. By permuting the objects one can form all $m!$ possible rankings. Considering the simplex $P_{m!}$, we wish to define a probability model, i.e. a family of probability distributions, i.e. a subset of $P_{m!}$, parametrized by θ in a space Θ: $\{P(\theta)|\theta \in \Theta\} \subset P_{m!}$, where $P(\theta)$ is a function from Θ to $P_{m!}$. The set of possible rankings of m objects has a group structure and is referred to as the symmetric group of order m, denoted $\mathbb{S}_m$. The distribution on $\mathbb{S}_m$ will be given through its density $P_\pi(\theta) = P[\Pi = \pi; \theta]$, $\pi \in \mathbb{S}_m, \theta \in \Theta$.

Please note that a *ranking* $\pi \in \mathbb{S}_m$ is a permutation of the object indices, i.e. indicating the ranks. Inverting a ranking gives the corresponding *ordering* $\varpi \in T_m$. An ordering lists the objects according to their order.

In *sufficient statistic models*, the parameter θ "touches the data π" only through functions $s(\pi)$. Section 9E of [8] motivate the *exponential family distributions*: if $s = (s_1, s_2, \ldots, s_p)$, then:

$$P_\theta(\pi^{(1)}, \ldots, \pi^{(n)}) = \exp(\sum_{j=1}^{p} \theta_j s_j - n\psi(\theta)).$$

We now look at an exponential family model using the pairwise comparisons $\mathrm{I}[\pi_a < \pi_b]$ implied by a ranking π as sufficient statistics. $\mathrm{I}[\pi_a < \pi_b]$ is the 0/1 indicator variable indicating whether the rank of object o_a is smaller than the rank of object o_b in the ranking π (meaning that object o_a is bigger/higher/more preferred/...). The model assumes that the structure resides in the pairwise comparisons. The general model is based on the $\binom{m}{2} \times 1$ parameter p whose indices ab, $a < b$ are ordered. The p_{ab} is interpreted as the probability object o_a would be preferred to object o_b if only that comparison were to be made. Note that $p_{ba} = 1 - p_{ab}$.

A ranking is obtained by making independently all the pairwise comparisons using those probabilities. The probability that the pairwise comparisons are consistent with an ordering $\varpi \in T_m$ is

$$Z(p) = \sum_{\varpi \in T_m} \prod_{a<b} p_{\varpi_a \varpi_b}$$

The probability of an ordering ϖ given that the pairwise comparisons are consistent is the probability that the comparisons yield ϖ divided by the probability they are consistent. The Babington Smith model [19] thus has the density

$$P_p(\varpi) = \frac{1}{Z(p)} \prod_{a<b} p_{\varpi_b \varpi_b}.$$

Remark: A Babington Smith model has weak stochastic transitivity, if for a, b,

$$p_{ab} \geq \frac{1}{2} \text{ and } p_{bc} \geq \frac{1}{2} \Rightarrow p_{ac} \geq \frac{1}{2}$$

and has strong stochastic transitivity if

$$p_{ab} \geq \frac{1}{2} \text{ and } p_{bc} \geq \frac{1}{2} \Rightarrow p_{ac} = \max\{p_{ab}, p_{bc}\}$$

We now write down the exponential model defined over the space of rankings, where the sufficient statistics consist of the $\bar{m} = \binom{m}{2}$ pairwise comparisons x_{ab} for $a < b$. The model is

$$\mathbf{M}(\pi|\theta) = \exp(\theta' \mathbf{X}(\pi) - \psi(\theta)), \quad \pi \in \mathbb{S}_m \tag{1}$$

where $\theta = (\theta_{12}, \theta_{13}, \ldots, \theta_{m-1,m})$,
$\mathbf{X}(\pi) = \mathbf{X}^\pi$ with $\mathbf{X} = (x_{12}, x_{13}, \ldots, x_{m-1,m})$,
$x_{ab}^\pi = I[\pi_a < \pi_b]$ (the pairwise comparisons implied by the ranking π),
$1 \leq a < b \leq m$;
ψ is the normalizing constant.

Note that the symmetric group (of rankings) is the model space, whereas the data space consists of all pairwise comparisons (matrices). The model enforces transitivities by comparing the measured, possibly intransitive choices with rank induced pairwise choices. Objects are ranked by determining the maximum likelihood ranking. Rankings with equal maximal likelihood are averaged.

The choice parameters p are related to the θ's through

$$p_{ab} = \frac{\exp(-\theta_{ab})}{1 + \exp(-\theta_{ab})} \quad , \; a, b \in \mathcal{O}. \tag{2}$$

The quantity $\mathbf{X}(\pi)$ plays the role of a dissimilarity measure. The model exemplifies the derivation of a suitable similarity information for non-Euclidean data that can be used in order to perform learning.

Given a sample of size n, the maximum likelihood estimator exists if and only if $0 < \hat{x}_{ab} < 1$ for all $a < b$. If $\hat{x}_{ab} = 0\ (= 1)$, then set $\hat{\theta}_{ab} = +\infty\ (-\infty)$. Let $H = \{(a, b) | a < b, 0 < \hat{x}_{ab} < 1\}$ be the set of pairs remaining, and $\mathbb{S}_m^*$ be the subset of rankings that conform to the sample, i.e.
$\mathbb{S}_m^* = \{\pi \in \mathbb{S}_m | \pi_a <\ (>)\ \pi_b \text{ if } \hat{x}_{ab} = 0\ (= 1)\}$.

The loglikelihood is

$$l^*(\theta, \hat{\mathbf{X}}) = \sum_{\substack{a<b \\ (a,b)\in H}} n\theta_{ab}\hat{x}_{ab} - n\psi^*(\theta) \tag{3}$$

with

$$\exp(\psi^*(\theta)) = \frac{1}{m!} \sum_{\pi \in \mathbb{S}_m^*} \exp\Big(\sum_{\substack{a<b \\ (a,b)\in H}} \theta_{ab} I[\pi_a < \pi_b] \Big) \tag{4}$$

3.1 Model-Based Clustering

For cluster analysis, the observed paired comparison data is assumed to consist of K groups. Each group is modeled by a Babington Smith distribution (cf. equation 1):

$$\mathbf{M}^{(k)}(\pi|\theta^{(k)}) = \exp(\theta^{(k)'}\mathbf{X}(\pi) - \psi(\theta^{(k)})), \quad \pi \in \mathbb{S}_m$$

The component distributions are joined in a mixture model,

$$\mathbf{M}(\pi) = \sum_{k=1}^{K} c^{(k)}\mathbf{M}^{(k)}(\pi|\theta^{(k)}), \tag{5}$$

with the mixture weights $(c^{(1)}, \ldots, c^{(K)})$ forming a partition of 1. Model parameters can be estimated with an expectation-maximization (EM) algorithm [20], or more sophisticated latent variable estimation algorithms such as Deterministic Annealing [11].

3.2 Missings

When measuring paired comparison data (e.g. elicit pairwise preferences in a survey), we have to expect that the pairwise comparison matrices may contain missings. That is, at position (a, b) in a matrix we do not have the information 0 or 1 but rather a $*$ indicating that this paired comparison is missing.

To further complicate the problem, in both cases below the pattern of missings might vary between the n pairwise comparison matrices constituting the samples.

Missings may occur for different reasons. The number of pairwise comparisons between m objects is $\frac{m(m-1)}{2}$. Instead of insisting on having all paired comparisons, the analyst might only measure/ask for a subset of the paired comparisons in order to make the experiment more cheap or comfortable. For example, he might query each pair with a probability p_q so that the number of necessary paired comparisons is only a fraction of all pairs.

A further reason for missings in a paired comparison dataset is that – though all paired comparisons are queried – some are not available. Some measurements might be unavailable, whether occuring by chance or built into the design of the experiment (e.g. to save costs or in an industrial experiment some results are missing because of mechanical breakdowns unrelated to the experimental process). Respondents in a survey might not answer all questions because they are indifferent w.r.t. to a paired comparison (i.e. object o_a and o_b are seen equally preferred; in an opinion survey some interviewees may be unable to express a preference for one object over another) or respondents get tired and are not willing to answer all the questions posed.

Sometimes it is natural to treat the values that are not observed as missing, in the sense that there are true underlying values that would have been observed if the industrial equipment had been better maintained or survey techniques had been better. Sometimes, however, it is less clear that a well-defined preference

has been masked by the nonresponse. Instead, the lack of a response is essentially an additional point in the sample space.

Excluding units that have missing values is generally inappropriate, since the investigator is usually interested in making inferences about the entire target population and since the removal of a subsample of common characteristic might cause a systematic bias.

In the following, an option for handling heterogeneous (i.e.: different patterns of missings within the samples) data in a probabilistic model is given. The performance of any missing-data method depends heavily on the mechanisms that lead to missing values. Data *missing completely at random* (MCAR) means that the missingness is not related to the data under study. Data can be missing at random (MAR, missingness is related to the observed data but not to the missing data) and there are also nonignorable missing-data mechanisms. For a thorough review of statistical analysis with missing data see the book of [16].

Notation:
$\text{Mis}^{(i)} = \{(a,b)|$ paired comparison between (a,b) missing in sample $i\}$
is the set of missings in sample i.

Model-based completion
Assuming that there are "true" values underlying at the missing matrix positions which are just masked (i.e. for each sample there is an unobservable complete pairwise comparison matrix), we can try to estimate these unobserved true values. We can explicitly estimate a maximum likelihood completion to a partial pairwise comparison matrix by treating the missing pairwise comparisons as latent information, and assuming complete pairwise comparison matrices to be distributed according to the model, e.g. the Babington-Smith model. An estimate of the full pairwise comparison matrix is obtained with an EM-type algorithm (latent variable estimation algorithm), which alternatingly reestimates the model parameters from current completion estimates, and then reestimates completions based on the current model (estimate the true frequencies of the full pairwise comparison matrices in the sample, then maximize the resulting likelihood).

In the E-step, the current parameter estimates are used to estimate the expected value of the sufficient statistics for the complete data. In the M-step, the estimated sufficient statistics are used to obtain maximum likelihood estimates of the model parameters.

This iterative EM procedure naturally suits into the clustering EM algorithm announced in section 3.1 and detailed in section 4. Having missings in the data adds more latent variables besides the unknown cluster assignments. The method can be used as basis for partial paired comparison data clustering, 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. Model-based completions can be performed based on the current cluster solution.

4 Model Inference

Heterogeneous, partial paired comparison data drawn from K distinct groups can now be described by a mixture model. The generative model for the data is

$$\mathbf{M}(\pi|c,\theta) = \sum_{k=1}^{K} c^{(k)} \frac{1}{m!} \exp(\theta^{(k)'} \mathbf{X}(\pi) - \psi^*(\theta^{(k)}, \mathrm{Mis}^{(\pi)})), \quad \pi \in \mathbb{S}_m \tag{6}$$

with the normalizing constant ψ depending on the cluster-specific $\theta^{(k)}$ and the sample-specific pattern of missings $\mathrm{Mis}^{(\pi)}$ by

$$\exp(\psi^*(\theta^{(k)}, \mathrm{Mis}^{(\pi)})) = \frac{1}{m!} \sum_{\rho \in \mathbb{S}_m^*} \exp(\sum_{\substack{a<b \\ (a,b)\in H}} \sum_{(a,b)\notin \mathrm{Mis}^{(\pi)}} \theta_{ab}^{(k)} I[\rho_a < \rho_b]).$$

For inference 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 [20]. In this section, we derive estimation equations for the heterogeneous data model, and discuss the implementation of an efficient EM algorithm for paired comparison data.

For data $\mathbf{X}^{(i)}$, $i = 1, \ldots, n$ and K clusters, define cluster assignments $q^{(i)} = (q^{(i)(1)}, \ldots, q^{(i)(K)})$. If $\mathbf{X}^{(i)}$ is assigned to cluster k, then $q^{(i)(k)} = 1$ and all other entries are 0. These assignment probabilities $q^{(i)(k)}$ ($q^{(i)(k)} \in [0,1]$, $\sum_k q^{(i)(k)} = 1$) are hidden variables of the EM estimation problem.

The E-step of the algorithm computes estimates of the assignment probabilities conditional on the current parameter configuration of the model. For samples that are only partially available, we want to make the cluster assignments maximally non-committal w.r.t. missings (i.e. paired comparisons not given). This involves establishing a uniform probability distribution over the missing values (maximum entropy argument), i.e. the restricted model assigns equal probabilities to all paired comparison matrices consistent with the given values regardless of what actual values the missings might have (uniform distribution over the missings). The maximum entropy approach avoids hidden assumptions about missing pairwise comparison entries. To summarize, for computing cluster assignments, the lack of knowledge about some paired comparisons is handled by substituting with the set of pairwise comparison matrices consistent with the given pairwise comparisons. The parameters θ are comparable for paired comparison matrices with different pattern of missings. Formally, this holds because the model is a distribution on the consistent completions (all possible matrices that are consistent with the given pairwise comparisons form an equivalence class).

Given estimates of the component parameter $\theta^{(k)}$ and the mixture weight $c^{(k)}$ for each cluster k, assignment probabilities are estimated as

$$q^{(i)(k)} = \frac{c^{(k)} \mathbf{M}(\pi^{(i)}|\theta^{(k)})}{\sum_{k'=1}^{K} c^{(k')} \mathbf{M}(\pi^{(i)}|\theta^{(k')})}.$$

In the M-step, assignment probabilities are assumed to be given. For each cluster, the parameters to be estimated are $c^{(k)}$ and $\theta^{(k)}$. As for any mixture model

EM algorithm, the mixture weights are straightforwardly computed as $c^{(k)} = \frac{1}{n}\sum_{i=1}^{n} q^{(i)(k)}$.

For ML estimation of the component parameters $\theta^{(k)}$, consider the Newton-Raphson method. To find the estimates of the $\theta_{ab}^{(k)}$ for each mode, the Newton-Raphson method can be applied to the negative log-likelihood:

$$-l^*(\theta^{(k)}, \hat{\mathbf{X}}) = \sum_{\substack{a<b \\ (a,b)\in H}} \sum_{i=1}^{n} q^{(i)(k)} \theta_{ab}^{(k)} \hat{x}_{ab}^{(i)} - \sum_{i=1}^{n} q^{(i)(k)} \psi^*(\theta^{(k)}) \tag{7}$$

In practice, the normalizing constants $\psi^*(\theta^{(k)}, \mathrm{Mis}^{(\pi)})$ can be expensive to compute if m is large. We therefore derived a MCMC sampler to approximate $l^*(\theta^{(k)})$ and $\psi^*(\theta^{(k)}, \mathrm{Mis}^{(\pi)})$.

Suppose that θ_0 is close to the ML estimate. A sample of rankings $\pi_{s_1}, \pi_{s_2}, \ldots, \pi_{s_s} \sim \mathbf{M}(\pi|\theta_0)$ is a random sample of rankings from the distribution defined by the paired comparison model with parameter θ_0. Make use of the law of large numbers to estimate an expectation (ML estimate in exponential families is the value $\hat{\theta}_0$ for which the expected value of the statistics is equal to the observed value) by a sample mean $\approx \frac{1}{s}\sum_{r=1}^{s} \exp((\theta - \theta_0)'\mathbf{X}(\pi_{s_r}))$ (further details of derivation omitted).

E-step: At the current parameter value $\theta^{(k)}$, a Monte Carlo simulation of the Markov ranking is made; this simulation is used to estimate cumulants (or moments) of the distribution. The approximated log-likelihood for cluster k is:

$$l^*(\theta^{(k)}, \hat{\mathbf{X}}) \approx \sum_{i=1}^{n} q^{(i)(k)} \Big\{ -\ln\{\frac{1}{s}\sum_{r=1}^{s} \exp((\theta^{(k)} - \theta_0^{(k)})'(\mathbf{X}(\pi_{s_r}^{(k)}) - \mathbf{X}(\pi^{(i)})))\}\Big\} \tag{8}$$

For sampling, simulate a discrete-time Markov chain whose stationary distribution is the distribution we want to sample from. Change (or not) the current ranking, according to some rule dependent on θ_0. Begining with an initial ranking, the elements of this ranking are stochastically updated, the updating mechanism circles through the ranking again and again, this defining a stochastic process which is a Markov chain. Approximate random draws e.g. by Gibbs sampling or Metropolis-Hastings.

Details for a Metropolis-Hastings type of sampler: As an elementary change, we define a transposition in the ranking, i.e. two random ranks are exchanged. π_τ denotes the ranking with transposition. The change takes effect with probability $\sim \min(1, p_\tau)$, with $p_\tau = \exp(\theta_0)'(\mathbf{X}(\pi_\tau) - \mathbf{X}(\pi))$, otherwise the change is discarded.

5 Selection of Comparisons

As pointed out in section 3.2, we should not rely on having all paired comparisons available, since the number of pairwise comparisons grows quadratically with

the number m of objects. Sometimes we have no control on *which* pairwise comparisons we can measure or get a response to. In other settings, however, we are able to *select* pairwise comparisons that data is aquired for. In a survey, for example, instead of asking for all pairwise comparisons, the interviewer can choose a subset of questions. It is thus reasonable to think of a strategy for the selection of comparisons.

We here again assume that there is a ranking underlying the paired comparisons (otherwise we see no argument why some paired comparisons are more "informative" than others). Under this transitivity assumption, the task reduces to the problem of *sorting* a partially ordered set (poset; the partial order induced by the paired comparisons). That is, like with any comparison-based sorting algorithm, one constructs a linear order (ranking) by queries "$<$" on pairs of objects. The two differences to standard sorting are: (i.) the query operation ("$<$") might be expensive (e.g. time-consuming measurement, limited attention of respondents in surveys); (ii.) the query operation ("$<$") might be noisy (e.g. flipped with a probability p_{Error}). We now give a method for selecting paired comparisons ensuring that the first paired comparisons queried are the most informative to construct the ranking. The method might not be robust to errors in the paired comparisons, in particular if errors occur between distant objects. For an error bound analysis for QuickSort with noisy comparison operation (resp. intransitivities) we refer to [1]. Probabilistic QuickSort always needs $O(n \log n)$ calls to the comparison oracle and, moreover, it is not clear whether the first queries yield the most valuable information about the ranking.

Let us try to make sure that the gain of information (for the ranking) is monotonically decreasing in the sequence of paired comparisons that are queried. The motivation is that only a limited number of comparisons can be queried due to cost or time constraints; for example, in a survey the interviewer does not even know when the interviewee will stop answering the questions. Technically, the problem rephrases as: Each additional comparison that is queried should reduce most the cardinality of the set of rankings (total orders) consistent with the partial order, since finally we would like to identify the single one underlying ranking. The method below is thus based on the theory of linear extensions [13].

Let P denote a poset (here: paired comparisons acquired so far) and $|E(P)|$ is the set of its linear extensions (here: all rankings consistent with the paired comparisons given). Suppose that we can choose any pair $o_a, o_b \in \{o_1, \ldots, o_m\}$ and ask an oracle to compare them. Having gotten the answer, say o_a precedes o_b, we add the relation $a < b$ and all its transitive consequences to P and obtain a new partial order $P^1 = P\&[a < b]$. We call the oracle again and ask it to compare a new pair of objects as long as $|E(P^q)| > 1$. In a finite number q of queries we sort the original poset P, i.e. obtain a total linear order $P^q = \pi \in E(P)$. Clearly, one has the information theory worst-case lower bound $q \geq \log_2 |E(P)|$ on the number of queries. For any poset P with $|E(P)| > 1$ linear extensions there exists a pair of objects $a, b \in \{1, \ldots, m\}$ such that:

$$\max\{\frac{|E(P\&[a < b])|}{|E(P)|}, \frac{|E(P\&[b < a])|}{|E(P)|}\} \leq \beta \tag{9}$$

The inequality (9) says that in any poset P there exists a comparison a, b which decreases the number of linear extensions by at least β. β is conjectured to be 2/3 and it is known [12] that inequality (9) holds with $\beta = 8/11$. The latter result implies that using 8/11-balanced test comparison one can sort an arbitrary poset P in at most $q \leq 2.2 \log_2 |E(P)|$ queries. [12] also show that computing the "balancing constants" $\beta_{ab} = |E(P\&[a < b])|/|E(P)| = \text{Prob}\{a < b \text{ in } E(P)\}$ is #P-hard. However, one can compute approximations to the balancing constants in time $O(T)$, where T is the complexity of nearly uniform generation of linear extensions of P. Therefore, a well-balanced comparison in a given poset can also be found with high probability in time $O(T)$. Now, the following fact [12]: Let $r_a = \frac{1}{|E(P)|} \sum_{\pi \in E} \pi(a)$ be the average rank of $a \in \{1, \ldots, m\}$ over the set of linear extensions of P, then an arbitrary pair a, b of objects such that $|r_a - r_b| < 1$ is an 8/11-balanced comparison in K. The strategy is to minimize $|\hat{r}_a - \hat{r}_b|$ over $a, b \in \{1, \ldots, m\}$. Intuitively, this approach favors comparing objects that are close to each other. This is particularly helpful to refine the underlying ranking, while it is – for exactly the same reason – of disadvantage for estimating a pairwise model, since comparisons between objects that are far apart (high absolute value of θ_{ab}) are more discriminative.

For averaging the ranks $\hat{r}_a$ of objects we need a nearly uniform generator of linear extensions of the poset (Markov chain with uniform stationary distribution for combinatorial object "linear extension"). For algorithm and details, please consult [13].

6 Application: Preference Prediction

Finally, we like to stress the usefulness of the probabilistic paired comparison model for preference modeling. A powerful approach to preference elicitation is the use of rankings, where the members of a population order items, values, or products according to their degree of preference or importance. The task of ranking, however, can be tedious. Deciding between just two items at a time is easier, and such pairwise preference data often naturally or implicitly arises (e.g. a dog is presented with two feeding dishes. The one that the dog eats first is the more preferred one).

Identifying Groups of Choices: The mixture model defined above expresses the separation of the comparators observed in the data into different groups or types, each of which exhibits a "typical" preference behavior. The interpretation of the θ_{ab}'s for each group is that a positive value codes a preference of object o_a over o_b by the group (when $\theta_{ab} \to \infty$: o_a is always preferred to o_b). A value of 0 means indifference or neutrality w.r.t. to the two objects at hand, whereas a negative value of θ_{ab} indicates that object o_a is seen as less preferable than o_b. The soft preference probabilities p_{ab} between objects can be used to construct the utility weights (as described in [22], for example) that a society and its groups assign to the different objects/options o_a, o_b.

Recommendations: The method is helpful to estimate preference relations on the set of objects, i.e. to predict the choice probabilities between two objects for

Table 1. Estimation errors on synthetic data with $K = 2$ and $K = 3$ clusters. The distances between the cluster centers and the cluster overlaps are varied.

Setting			Results
K	τ_d	λ / p	$\text{error}_{L2}\ \hat{\theta}$
2	6	1.0 / 0.73	0.43 ± 0.10
	3	1.0 / 0.73	0.55 ± 0.12
		0.5 / 0.62	0.52 ± 0.17
			l-approximation: 0.94 ± 0.60
3	4-5	1.5 / 0.82	1.14 ± 0.42
	2-4	1.5 / 0.82	0.84 ± 0.27
		0.5 / 0.62	0.74 ± 0.13

an individual. The prediction of the preference between objects o_a and o_b for individual i based on the cluster solution is given by the posterior:

$$p_{ab}^{(i)} = \sum_{k=1}^{K} q^{(i)(k)} \frac{\exp(-\theta_{ab}^{(k)})}{1 + \exp(-\theta_{ab}^{(k)})} \tag{10}$$

7 Experimental Results

The experiments include synthetic and real-world paired comparison data derived from rankings. The mixture analysis with artificial data drawn from a density with known parameters is conducted to check the method's capability of recovering parameters from paired comparison data. Additional experiments are conducted on a data set from a study on change in mass politics described in [3], where Germans expressed their preference regarding political goals. All experiments are performed with the EM algorithm described in section 4.

7.1 Synthetic Data

Synthetic pairwise comparison data observations were derived from rankings drawn at random from a mixture of Mallows models [18]. Sample experiments for $m = 4$ objects and $K = 2$ and 3 clusters are shown in Tab 1. τ_d are the mutual Kendall distances between the cluster centers; λ is an inverse spread (a lower value resulting in an higher overlap between the clusters) and $1 - p$ is the inverse flip probability of a pairwise comparison. In the setting of 2 equally sized clusters $n = 150$ samples were used, for the 3 clusters, including a small one, $n = 300$ samples were used. The quality of parameter estimates is reported as the L2 error to the true θ. The Bayesian Information Criterion (BIC) [20] estimate of the number of clusters was correct except for very near cluster centers and/or broad cluster overlaps.

With the distance between the cluster centers decreasing, the estimation errors increase. The estimation errors become smaller when the cluster have a higher spread (small λ). Approximating the likelihood by sampling generally increases the estimation error.

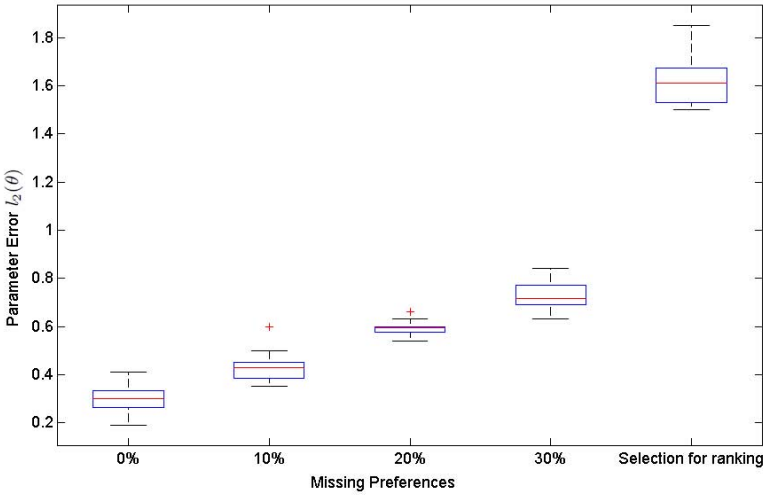

Fig. 1. Missings in the data set: Estimation error of θ versus different sorts of incomplete data

We also measured the parameter recovery error depending on different types of missings in the data, i.e. with the paired comparison matrices only partially available in the sample (here: $n = 500$). The value of the method being capable of handling missings is illustrated in Fig. 1, where from left to right the number of missings increases (Remark: thereby, the effective sample size is reduced, may possibly reduce e.g. the costs of measuring or time of surveying!). First, the algorithm sees all pairwise comparisons (the full information is available). In the second scenario, in each pairwise comparison matrix, each entry is available with probability 0.9. It is a genuine advantage of the proposed model that it can handle samples containing different patterns of missing at the same time. Previously, when the data analyst was confronted with such heterogenous data it was often common practice that he had to delete incomplete samples or to analyse them separately. Next, random 80% and 70% of the pairwise comparisons got accessible to the inference procedure. Finally, we used the method described in section 5 to automatically determine the subset of paired comparisons for ranking construction. As expected, the error is significantly higher compared to random selection for the reason given at the end of section 5: comparisons between near objects are helpful to refine the underlying ranking, while for discriminating between clusters distant objects are more helpful.

7.2 Political Goals German Data

The political goals data set of real-world rankings from a study on change in mass politics: A sample of 2262 Germans expressed their preference on four political goals based on their perceived personal importance: *Order*, *Say*, *Prices*, *Freedom*. We analyzed the paired comparisons by EM estimation of the above mixture

Table 2. Political Goals: Preference probabilities for the two clusters found: "Materialism" and "Post-materialism"

Pair $(o_a, o_b) \rightarrow$	(O,S)	(O,P)	(O,F)	(S,P)	(S,F)	(P,F)
$\hat{p}_{ab}^{(Mat.)}$	77%	48%	86%	29%	53%	76%
$\hat{p}_{ab}^{(Post-mat.)}$	43%	37%	29%	58%	57%	49%

model and found two clusters in agreement with the original classification by [3] of the goals into *Materialist* and *Post-materialist* (see Tab. 2). The analysis in [19] by a simple Babington Smith model "lefts a significant proportion of the data unexplained". We measured the prediction quality of our method by deleting 10% random subsamples of the paired comparisons. The trained model was able to predict the capped paired choice probabilities with a prediction error of 8.65% $\pm$ 0.78%. To the best of our knowledge, there does not exist an alternative method for comparison that is able to make predictions on this granularity of individual paired choices.

8 Conclusion

A probabilistic mixture model for the analysis of inhomogeneous paired comparison data was introduced. Our modeling approach permits the integration of data with different patterns of missings by estimating a model-based distribution on the subset of matrices consistent with the information given and thus can combine estimate contributions in a meaningful way.

The assumption throughout this line of work is that there is a ranking underlying the order relation. A ranking (or total order) orders objects according to some criterion, neglecting any "distance" between the objects. In practice, paired comparisons (or partial orders) are sometimes easier to acquire. In fact, when rankings are distributed according to the well-known Mallows model with modal ranking σ and inverse spread λ, the flip probabilities of the induced paired comparisons directly relate to the spread of the rank model. An advantage of models based on ranks is that parameters can be tied in order to reduce the number of free parameters (see [9,4]).

The underlying ranking assumption is valid as long as there is a single criterion under which the objects are evaluated, or the objects map to a linear scale. What can be done in case of intransitivities ($o_a < o_b$, $o_b < o_c$, and $o_c < o_a$) that arise systematically due to conflicting multiple criteria? Intransitivities can be consistently resolved and used to estimate utility weights for multicriteria decision making ([5]).

References

1. Ailon, N., Mohri, M.: An Efficient Reduction of Ranking to Classification, Technical Report, TR2007-903 (2007)
2. Ailon, N., Mohri, M.: Preference-Based Learning to Rank. Machine Learning 80, 189–211 (2010)

3. Barnes, S.H., Kaase, M.: Political Action: Mass Participation in Five Western Countries. Sage, Beverly Hills (1979)
4. Busse, L.M., Orbanz, P., Buhmann, J.M.: Cluster Analysis of Heterogeneous Rank Data. In: International Conference on Machine Learning (2007)
5. Busse, L.M., Buhmann, J.M.: Multicriteria Scaling for Utilities under Intransitivities (to appear, 2011)
6. Cao, Z., Qin, T., Liu, T.-Y., Tsai, M.-F., Li, H.: Learning to Rank: From Pairwise Approach to Listwise Approach, Microsoft Tech. Report (2007)
7. Cohen, W.W., Schapire, R.E., Singer, Y.: Learning to Order Things. In: Advances in Neural Information Processing Systems, vol. 10 (1998)
8. Diaconis, P.: Group Representations in Probability and Statistics, Institute of Mathematical Statistics (1988)
9. Fligner, M.A., Verducci, J.S.: Distance based rank models. Journal of the Royal Statistical Society B 48(3), 359–369 (1986)
10. Fürnkranz, J., Hüllermeier, E.: Pairwise Preference Learning and Ranking. In: Lavrač, N., Gamberger, D., Todorovski, L., Blockeel, H. (eds.) ECML 2003. LNCS (LNAI), vol. 2837, pp. 145–156. Springer, Heidelberg (2003)
11. Hofmann, T., Buhmann, J.: Pairwise Data Clustering by Deterministic Annealing. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(1), 1–14 (1997)
12. Kahn, J., Saks, M.: Every poset has a good comparison. In: Proc. 16-th Symposium on Theory of Computing, pp. 299–301 (1984)
13. Karzanov, A., Khachiyan, L.: On the Conductance of Order Markov Chains. Order 8, 7–15 (1991)
14. Kendall, M.G., Babington Smith, B.: On the Method of Paired Comparisons. Biometrika 31, 324–345 (1940)
15. Lebanon, G., Lafferty, J.D.: Cranking: Combining Rankings Using Conditional Probability Models on Permutations. In: International Conference on Machine Learning (2002)
16. Little, R.J.A., Rubin, D.B.: Statistical analysis with missing data. Wiley series in probability and mathematical statistics. Applied probability and statistics, NJ (2002)
17. Lu, T., Boutilier, C.: Learning Mallows Models with Pairwise Preferences. In: International Conference on Machine Learning (2011)
18. Mallows, C.L.: Non-null ranking models I. Biometrika 44, 114–130 (1957)
19. Marden, J.I.: Analyzing and Modeling Rank Data. Chapman & Hall, Boca Raton (1995)
20. McLachlan, G.J., Krishnan, T.: The EM Algorithm and Extensions. John Wiley & Sons, Chichester (1997)
21. Meila, M., Phadnis, K., Patterson, A., Bilmes, J.: Consensus ranking under the exponential model. In: Conference on Uncertainty in Artificial Intelligence, UAI (2007)
22. Saaty, T.L.: A scaling method for priorities in hierarchical structures. Journal of Mathematical Psychology 15, 234–281 (1977)
23. Su, X., Khoshgoftaar, T.M.: A Survey of Collaborative Filtering Techniques. In: Advances in Artificial Intelligence (2009)