MRA-BASED STATISTICAL LEARNING FROM INCOMPLETE RANKINGS

Eric Sibony    Stéphan Clémençon    Jérémie Jakubowicz
ICML, July 7 2015

LTCI UMR 5141, Telecom ParisTech/CNRS
Ranking data naturally appear in a wide variety of situations

- elections
- survey answers
- expert judgments
- race results
- competition rankings
- customers behaviors
- users preferences
- ...
Ranking Data

- Catalog of items $\mathbb{J} := \{1, \ldots, n\}$
- Each observation is a partial order $\prec$ on $\mathbb{J}$

Example

- Full ranking: $a_1 \succ \cdots \succ a_n$
- Top-k ranking: $a_1 \succ \cdots \succ a_k \succ$ the rest
- Pairwise comparison: $a \succ b$
Probabilistic modeling

Full ranking $a_1 \prec \cdots \prec a_n$

$\iff$ Permutation $\sigma \in \mathcal{S}_n$ that maps an item to its rank: $\sigma(a_i) = i$

The variability of full rankings is therefore modeled by a probability distribution $p$ over the set of permutations $\mathcal{S}_n$. $p$ is called a ranking model.
Probabilistic modeling

For a general ranking $≺$, let $S$ be the set of its linear extensions:

$$S = \{ \sigma \mid a ≻ b \Rightarrow \sigma(a) < \sigma(b) \} \subset \mathcal{S}_n$$

The observation of $≺$ is then modeled as a censored observation of a latent random permutation $\Sigma \sim p$:

$$P[≺] = P[\Sigma \in S] = \sum_{\sigma \in S} p(\sigma)$$

Example

$$P[a ≻ \text{the rest}] = P[\Sigma(a) = 1] = \sum_{\sigma \in \mathcal{S}_n, \sigma(a) = 1} p(\sigma)$$

$$P[a ≻ b] = P[\Sigma(a) < \Sigma(b)] = \sum_{\sigma \in \mathcal{S}_n, \sigma(a) < \sigma(b)} p(\sigma)$$
Only some marginals of the ranking model $p$ are observed.

**Crucial difference: partial vs incomplete rankings**

- **Partial rankings** provide *absolute* rank information:

  $$a_{1,1} \cdots a_{1,n_1} \succ \cdots \succ a_{r,1}, \ldots, a_{r,n_r}$$

  with $\sum_{i=1}^{r} n_i = n$

  (all the items are involved in each ranking)

- **Incomplete rankings** provide *relative* rank information:

  $$a_1 \succ \cdots \succ a_k$$

  with $k < n$

  (only a subset of items are involved in each ranking)
Many existing approaches naturally apply to partial rankings

Mallows model ([Mallows, 1957], [Lebanon and Mao, 2008]), Plackett-Luce model ([Plackett, 1975]), Metric methods ([Critchlow, 1985]), Harmonic analysis ([Diaconis, 1988], [Kakarala, 2012]), Riffled independence ([Huang et al., 2012]), …

Much fewer apply to incomplete rankings

- Plackett-Luce model
- Mallows model with the methods introduced in [Lu and Boutilier, 2011]
- The kernel-based methods introduced in [Kondor and Barbosa, 2010] and [Sun et al., 2012]
We focus on incomplete rankings and consider the problem of the estimation of the accessible marginals.

Our contributions are threefold:

1. We define a rigorous setting for the statistical estimation of marginals associated to incomplete rankings.

2. We exploit the multiresolution analysis constructed in [Clémençon et al., 2014] to build an estimator.

3. We prove theoretical guarantees about its accuracy and complexity.
STATISTICAL SETTING
A full ranking $\sigma$ extends an incomplete ranking $\pi$ if the items involved in $\pi$ are ranked in the same order in $\sigma$.

**Example for** $n = 5$

$$\sigma := 2 \gg 5 \gg 4 \gg 1 \gg 3$$ extends $$\pi := 5 \gg 1 \gg 3.$$

We denote by $\pi \subset \sigma$. 
For any finite set $E$, we define $\mathcal{P}(E) := \{A \subset E | |A| \geq 2\}$

**Definition (Marginal)**

The marginal of the ranking model $p$ on a subset of items $A \in \mathcal{P}([n])$ is the probability distribution over the rankings that involve the items of $A$ defined by

$$P_A(\pi) = \sum_{\sigma \in S_n, \pi \subset \sigma} p(\sigma)$$

**Example for $n = 4$**

$$P_{\{1,3,4\}}(4 \succ 1 \succ 3) = p(2 \succ 4 \succ 1 \succ 3) + p(4 \succ 2 \succ 1 \succ 3) + p(4 \succ 1 \succ 2 \succ 3) + p(4 \succ 1 \succ 3 \succ 2)$$
Notations

\( \Gamma(A) = \{\text{rankings that involve the items of } A\} \text{ for } A \in \mathcal{P}([n]) \)

\( (\Gamma([n])) \Leftrightarrow \mathcal{S}_n \)

\( \Gamma_n = \bigsqcup_{A \in \mathcal{P}([n])} \Gamma(A) = \{\text{all incomplete rankings on } [n]\} \)

Definition (Marginal operator)

For \( A \in \mathcal{P}([n]) \), define the operator

\[ M_A : \mathbb{R}^{\Gamma_n} \rightarrow \mathbb{R}^{\Gamma(A)}, \quad f \mapsto M_A f \]

by

\[ M_A f(\pi) = \sum_{\sigma \in \Gamma_n, \pi \subset \sigma} f(\sigma) \text{ for } \pi \in \Gamma(A) \]
Observations: $1 \succ 2$, $3 \succ 1 \succ 4$, $2 \succ 5$, $5 \succ 3$, $2 \succ 1 \succ 5$, $4 \succ 3 \succ 5$, $4 \succ 1 \succ 3 \succ 2$, $3 \succ 5$, $1 \succ 2$, $\ldots$

Two types of variability:

- Variability of the observed subset of items $A \in \mathcal{P}([n])$
- Variability of the observed ranking $\pi$ among $\Gamma(A)$

**Definition (Observation process)**

Each observation is modeled as a couple $(A, \Pi)$ drawn from the following scheme

$$A \sim \nu \quad \text{and} \quad \Pi|\{A = A\} \sim P_A$$

where $\nu$ is a probability distribution on $\mathcal{P}([n])$

Dataset: $(A_1, \Pi^{(1)}), \ldots, (A_N, \Pi^{(N)})$ IID
The information from p is censored through $\nu$, more specifically through its support $\mathcal{A} = \{A \in \mathcal{P}([n]) \mid \nu(A) > 0\}$

Without any structural assumption on $p$, only the parameters from the marginals $P_A$ for $A \in \mathcal{A}$ are accessible

**Example**

If one only observes pairwise comparisons, then she only has access to marginals $P_{\{a,b\}}$ for distinct $a, b \in [n]$

$\Rightarrow$ Leads to at most $O(n^2)$ parameters, $\ll n! - 1$ parameters of $p$
Construct an empirical ranking model $\hat{q}_N : \mathcal{S}_n \to \mathbb{R}$ from the dataset such that its marginals $M_A \hat{q}_N$ are good estimators of the $P_A$’s for $A \in \mathcal{A}$.

Error measure

The empirical ranking model $\hat{q}_N$ is evaluated through

$$\mathcal{E}(\hat{q}_N) = \mathbb{E} \left[ \sum_{A \in \mathcal{A}} \nu(A) \| M_A \hat{q}_N - P_A \|^2_A \right]$$

where the expectation is over the drawings of the dataset and $\| \cdot \|_A$ is the euclidean norm on $\mathbb{R}^{\Gamma(A)}$. 
One can build from the dataset the empirical estimator for each observed subset $A$

$$\hat{P}_A(\pi) = \frac{|\{1 \leq i \leq N \mid \Pi^{(i)} = \pi\}|}{|\{1 \leq i \leq N \mid A_i = A\}|}$$

for $\pi \in \Gamma(A)$
(\{1,3\}, 13), (\{1,2,3\}, 231), (\{1,3,4\}, 143), (\{1,3\}, 31), (\{2,4\}, 24), (\{2,4\}, 42), (\{3,4\}, 43), (\{1,2,3\}, 213), (\{3,4\}, 43), (\{1,3,4\}, 413), (\{1,3,4\}, 134), ...
One can build from the dataset the empirical estimator for each observed subset $A$

$$\hat{P}_A(\pi) = \frac{|\{1 \leq i \leq N \mid \Pi^{(i)} = \pi\}|}{|\{1 \leq i \leq N \mid A_i = A\}|}$$

for $\pi \in \Gamma(A)$
One can build from the dataset the empirical estimator for each observed subset $A$

$$\hat{P}_A(\pi) = \frac{|\{1 \leq i \leq N \mid \Pi^{(i)} = \pi\}|}{|\{1 \leq i \leq N \mid A_i = A\}|} \text{ for } \pi \in \Gamma(A)$$

**Issues**

- There may not exist a ranking model $\hat{q}_N$ such that $M_A\hat{q}_N = \hat{P}_A$ for $A \in \mathcal{A}$
- The information inferred from the observations is not consolidated
Idea

Look for an empirical ranking model $\hat{q}_N$ such that

$$M_A \hat{q}_N \approx \hat{P}_A$$

for each observed $A$

Challenges

Need to deal with a linear system

- Of very high dimensionality ($n! - 1$ parameters)
- Where the equations are intertwined through complex combinatorial relationships
MRA FRAMEWORK
To exploit the multiscale structure of marginals.
Example for $n = 4$: 

\[
\begin{align*}
F &\quad M_{\{1,2,3\}} F \\
&\quad M_{\{1,2,4\}} F \\
&\quad M_{\{1,3,4\}} F \\
&\quad M_{\{2,3,4\}} F \\
\end{align*}
\]
To exploit the **multiscale structure of marginals**.

Example for $n = 4$:

$$
\begin{align*}
& \Rightarrow \text{Construct features that localize the part of information that is specific to each marginal}
\end{align*}
$$
Theorem ([Clémençon et al., 2014])

For any $F : \Gamma(A) \to \mathbb{R}$,

$$F = \phi_A \sum_{B \in \mathcal{P}(A) \cup \{\emptyset\}} X_B F$$

In addition for all $A' \in \mathcal{P}(A)$,

$$M_{A'}F = \phi_{A'} \sum_{B \in \mathcal{P}(A') \cup \{\emptyset\}} X_B F$$

where

- $X_B F$: wavelet projection, vector in a features space $\mathbb{H}_n$
- $\phi_A$: embedding operator from the features space to the signal space $\mathbb{R}^{\Gamma(A)}$
Example

For a function $F$

$$M_{\{1,2,3\}}F = \phi_{\{1,2,3\}} \left[ X_{\emptyset}F + X_{\{1,2\}}F + X_{\{1,3\}}F + X_{\{2,3\}}F + X_{\{1,2,3\}}F \right]$$

For the ranking model $p$

$$P(2 \succ 1 \succ 3) = \frac{1}{6} + \frac{1}{2} \left[ \left( P(2 \succ 1) - \frac{1}{2} \right) + \left( P(1 \succ 3) - \frac{1}{2} \right) \right] + \text{residual}$$
\[ \phi_A \cdot X_0 F = \frac{1}{|A|!} \sum_{\pi \in \Gamma_n} F(\pi) \]

- Level 2 information specific to \( \{a, b\} \)

\[ X_{\{a, b\}} F = M_{\{a, b\}} F - \phi_{\{a, b\}} X_0 F \]

- Level 3 information specific to \( \{a, b, c\} \)

\[ X_{\{a, b, c\}} F = M_{\{a, b, c\}} F - \phi_{\{a, b, c\}} [X_{\{a, b\}} F + X_{\{a, c\}} F + X_{\{b, c\}} F + X_0 F] \]
Signal space

Features space

Analyis

Synthesis

\[ \mathbb{R}^\Gamma(A) \]

\[ \mathbb{R}^\Gamma(A') \]

\[ F \]

\[ M_{A'} \]

\[ M_{A'}F \]

\[ \phi_A \]

\[ \phi_{A'} \]

\[ (X_B F)_{B \in \mathcal{P}(A) \cup \{\emptyset\}} \]

\[ (X_B F)_{B \in \mathcal{P}(A') \cup \{\emptyset\}} \]

\[ \mathbb{H}_n \]

\[ \mathbb{H}_n \]
General method

1. Construct estimators $\hat{X}_B$ of $X_{B^p}$ for subsets $B$ in a collection $\mathcal{B} \subset \mathcal{P}([n])$

2. The empirical ranking model is defined by

$$\hat{q}_N = \phi_{[n]} \sum_{B \in \mathcal{B}} \hat{X}_B$$

The estimator of a marginal on a subset $A \in \mathcal{P}([n])$ is then given by

$$M_A \hat{q}_N = \phi_A \sum_{B \in \mathcal{B} \cap \mathcal{P}(A)} \hat{X}_B$$
Denote by $N_A$ the number of times the subset $A \in \mathcal{P}(\mathbb{[n]})$ was observed, so that $\sum_{A \in \mathcal{A}} N_A = N$.

The WLS estimator

1. Construct the empirical estimators $\hat{P}_A$ for each observed subset $A$
2. Compute their wavelet projections $X_B \hat{P}_A$ for $B \in \mathcal{P}(A)$
3. For each $B \in \mathcal{P}(\mathbb{[n]})$, the WLS estimator is then defined as the weighted average of the $X_B \hat{P}_A$'s for $A \supset B$:

$$\hat{X}_B^{WLS} = \sum_{A \supset B} \frac{N_A}{\sum_{A' \supset B} N_A'} X_B \hat{P}_A$$
Illustration
PROPERTIES OF THE MRA-BASED ESTIMATOR
We recall that $N$ is the size of the dataset

**Theorem**

1. The WLS estimator is asymptotically unbiased:

$$\lim_{N \to \infty} \mathbb{E} \left[ M_A \hat{q}_N^{WLS} \right] = P_A \quad \text{for each } A \in \mathcal{A}$$

2. The error of the WLS estimator is bounded by

$$\mathcal{E}(\hat{q}_N^{WLS}) \leq \frac{C_1}{N} + C_2 \rho^{2N}$$

where $C_1$ and $C_2$ are constants that only depend on $p$ and $\nu$ (in particular not directly on $n$)
Set $K = \max_{A \in \mathcal{A}} |A|$

**Theorem**

1. The learning complexity is bounded by

   $$(K! + 1)2^K(|\mathcal{A}| + N)$$

2. The number of storage parameters is bounded by

   $$K! 2^K \min(|\mathcal{A}|, N)$$

3. The computation of each $M_A q_N^{WLS}(\pi)$ is bounded by

   $$\frac{|A|(|A| - 1)}{2}$$
Comparison of 4 empirical ranking models:

- The WLS estimator
- The collection of empirical estimators of the marginals
- The Plackett-Luce model fitted with the MM algorithm from [Hunter, 2004]
- A state-of-the-art estimator from [Sun et al., 2012]

On the APA Dataset, for $\nu$ uniform up to scale $K \in \{2, \ldots, 5\}$. 
ILLUSTRATION ON NUMERICAL EXPERIMENTS

Up to scale 2

Up to scale 3

Up to scale 4

Up to scale 5
FUTURE DIRECTIONS

- Application to other statistical tasks (e.g. prediction)
- Definition of efficient regularization procedures for large values of n
- Extension to incomplete rankings with ties


Learning mallows models with pairwise preferences.
In ICML, pages 145–152.

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

The analysis of permutations.

Estimating probabilities in recommendation systems.