Ranking Median Regression: Learning to Order through Local Consensus

Anna Korba[⋆] Stéphan Clémençon[⋆] Eric Sibony[†]

⋆ Telecom ParisTech, † Shift Technology

Statistics/Learning at Paris-Saclay @IHES
January 19 2018

Outline

1. Introduction to Ranking Data
2. Background on Ranking Aggregation
3. Ranking Median Regression
4. Local Consensus Methods for Ranking Median Regression
5. Conclusion

Outline

Introduction to Ranking Data

Background on Ranking Aggregation

Ranking Median Regression

Local Consensus Methods for Ranking Median Regression

Conclusion

Ranking Data

Set of items $[\![n]\!] := \{1, \ldots, n\}$

Definition (Ranking)

A ranking is a strict partial order $\prec$ over $[\![n]\!]$, *i.e.* a binary relation satisfying the following properties:

Irreflexivity For all $i \in [\![n]\!]$, $i \not\prec i$

Transitivity **For all i, j, k $\in [\![n]\!]$, if i $\prec$ j and j $\prec$ k then i $\prec$ k**

Asymmetry For all $i, j \in [\![n]\!]$, if $i \prec j$ then $j \not\prec i$

Ranking data arise in a lot of applications

Traditional applications

- **Elections**: $[\![n]\!]$= a set of candidates
  $\rightarrow$ A voter ranks a set of candidates
- **Competitions**: $[\![n]\!]$= a set of players
  $\rightarrow$ Results of a race
- **Surveys**: $[\![n]\!]$= political goals
  $\rightarrow$ A citizen ranks according to its priorities

Modern applications

- **E-commerce**: $[\![n]\!]$= items of a catalog
  $\rightarrow$ A user expresses its preferences (see ”implicit feedback”)
- **Search engines**: $[\![n]\!]$= web-pages
  $\rightarrow$ A search engine ranks by relevance for a given query

The analysis of ranking data spreads over many fields of the scientific literature

- Social choice theory
- Economics
- Operational Research
- Machine learning

$\Rightarrow$ Over the past 15 years, the statistical analysis of ranking data has become a subfield of the machine learning literature.

Many efforts to bring them together

NIPS 2001	New Methods for Preference Elicitation
NIPS 2002	Beyond Classification and Regression
NIPS 2004	Learning with Structured Outputs
NIPS 2005	Learning to Rank
IJCAI 2005	Advances in Preference Handling
SIGIR 07-10	Learning to Rank for Information Retrieval
ECML/PKDD 08-10	Preference Learning
NIPS 09	Advances in Ranking
NIPS 2011	Choice Models and Preference Learning
EURO 09-16	Special track on Preference Learning
ECAI 2012	Preference Learning
DA2PL 2012,2014,2016	From Decision Analysis to Preference Learning
Dagstuhl 2014	Seminar on Preference Learning
NIPS 2014	Analysis of Rank Data
ICML 2015-2017	Special track on Ranking and Preferences
NIPS 2017	Learning on Functions, Graphs and Groups

Common types of rankings

Set of items $[\![n]\!] := \{1, \dots, n\}$

- **Full ranking.** All the items are ranked, without ties

$$a_1 \succ a_2 \succ \cdots \succ a_n$$

- **Partial ranking.** All the items are ranked, with ties (”buckets”)

$$a_{1,1}, \dots, a_{1,n_1} \succ \cdots \succ a_{r,1}, \dots, a_{r,n_r} \quad \text{with} \quad \sum_{i=1}^{r} n_i = n$$

$\Rightarrow$ **Top-k ranking** is a particular case: $a_1, \dots, a_k \succ$ the rest

- **Incomplete ranking.** Only a subset of items are ranked, without ties

$$a_1 \succ \cdots \succ a_k \quad \text{with} \quad k < n$$

$\Rightarrow$ **Pairwise comparison** is a particular case: $a_1 \succ a_2$

Detailed example: analysis of full rankings

Notation.

- A full ranking: $a_1 \succ a_2 \succ \cdots \succ a_n$
- Also seen as the permutation σ that maps an item to its rank:

$$a_1 \succ \cdots \succ a_n \quad \Leftrightarrow \quad \sigma \in \mathfrak{S}_n \text{ such that } \sigma(a_i) = i$$

$\mathfrak{S}_n$: set of permutations of $[\![n]\!]$, the symmetric group.

Probabilistic Modeling. The dataset is a collection of random permutations drawn IID from a probability distribution P over $\mathfrak{S}_n$:

$$\mathcal{D}_N = (\Sigma_1, \ldots, \Sigma_N) \quad \text{with} \quad \Sigma_i \sim P$$

P is called a ranking model.

Detailed example: analysis of full rankings

- Ranking data are very natural for human beings
  $\Rightarrow$ Statistical modeling should capture some interpretable structure

Questions

- How to analyze a dataset of permutations $\mathcal{D}_N = (\Sigma_1, \ldots, \Sigma_N)$?
- How to characterize the variability? What can be inferred?

Detailed example: analysis of full rankings

Challenges

- A random permutation Σ can be seen as a random vector $(\Sigma(1), \ldots, \Sigma(n)) \in \mathbb{R}^n$... but

Detailed example: analysis of full rankings

Challenges

- A random permutation Σ can be seen as a random vector $(\Sigma(1), \ldots, \Sigma(n)) \in \mathbb{R}^n$... but
  The random variables $\Sigma(1), \ldots, \Sigma(n)$ are highly dependent and the sum $\Sigma + \Sigma'$ is not a random permutation!
  $\Rightarrow$No natural notion of variance for Σ

Detailed example: analysis of full rankings

Challenges

- A random permutation Σ can be seen as a random vector $(\Sigma(1), \ldots, \Sigma(n)) \in \mathbb{R}^n$... but
  The random variables $\Sigma(1), \ldots, \Sigma(n)$ are highly dependent and the sum $\Sigma + \Sigma'$ is not a random permutation!
  $\Rightarrow$No natural notion of variance for Σ
- The set of permutations $\mathfrak{S}_n$ is finite... but

Detailed example: analysis of full rankings

Challenges

- A random permutation Σ can be seen as a random vector $(\Sigma(1), \ldots, \Sigma(n)) \in \mathbb{R}^n$... but
  The random variables $\Sigma(1), \ldots, \Sigma(n)$ are highly dependent and the sum $\Sigma + \Sigma'$ is not a random permutation!
  $\Rightarrow$No natural notion of variance for Σ
- The set of permutations $\mathfrak{S}_n$ is finite... but
  Exploding cardinality: $|\mathfrak{S}_n| = n!$
  $\Rightarrow$ Few statistical relevance

Detailed example: analysis of full rankings

Challenges

- A random permutation Σ can be seen as a random vector $(\Sigma(1), \ldots, \Sigma(n)) \in \mathbb{R}^n$... but
  The random variables $\Sigma(1), \ldots, \Sigma(n)$ are highly dependent and the sum $\Sigma + \Sigma'$ is not a random permutation!
  $\Rightarrow$No natural notion of variance for Σ
- The set of permutations $\mathfrak{S}_n$ is finite... but
  Exploding cardinality: $|\mathfrak{S}_n| = n!$
  $\Rightarrow$ Few statistical relevance
- Apply a method from p.d.f. estimation (e.g. kernel density estimation)... but

Detailed example: analysis of full rankings

Challenges

- A random permutation Σ can be seen as a random vector $(\Sigma(1), \ldots, \Sigma(n)) \in \mathbb{R}^n$... but
  The random variables $\Sigma(1), \ldots, \Sigma(n)$ are highly dependent and the sum $\Sigma + \Sigma'$ is not a random permutation!
  $\Rightarrow$No natural notion of variance for Σ
- The set of permutations $\mathfrak{S}_n$ is finite... but
  Exploding cardinality: $|\mathfrak{S}_n| = n!$
  $\Rightarrow$ Few statistical relevance
- Apply a method from p.d.f. estimation (e.g. kernel density estimation)... but
  No canonical ordering of the rankings!

Main approaches

"Parametric" approach

- Fit a predefined generative model on the data
- Analyze the data through that model
- Infer knowledge with respect to that model

"Nonparametric" approach

- Choose a structure on $\mathfrak{S}_n$
- Analyze the data with respect to that structure
- Infer knowledge through a "regularity" assumption

Parametric Approach - Classic Models

- Thurstone model [Thurstone, 1927]
  Let $\{X_1, X_2, \ldots, X_n\}$ r.v with a continuous joint distribution $F(x_1, \ldots, x_n)$:

$$P(\sigma) = \mathbb{P}(X_{\sigma^{-1}(1)} < X_{\sigma^{-1}(2)} < \cdots < X_{\sigma^{-1}(n)})$$

- Plackett-Luce model [Luce, 1959], [Plackett, 1975]
  Each item i is parameterized by w_i with $w_i \in \mathbb{R}^+$:

$$P(\sigma) = \prod_{i=1}^{n} \frac{w_{\sigma_i}}{\sum_{j=i}^{n} w_{\sigma_j}}$$

  Ex: $2 \succ 1 \succ 3 = \frac{w_2}{w_1+w_2+w_3} \frac{w_1}{w_1+w_3}$

- Mallows model [Mallows, 1957]
  Parameterized by a central ranking $\sigma_0 \in \mathfrak{S}_n$ and a dispersion parameter $\gamma \in \mathbb{R}^+$

$$P(\sigma) = Ce^{-\gamma d(\sigma_0, \sigma)} \qquad \text{with } d \text{ a distance on } \mathfrak{S}_n.$$

Nonparametric approaches - Examples 1

- Embeddings
  - Permutation matrices [Plis et al., 2011]

$$\mathfrak{S}_n \to \mathbb{R}^{n\times n}, \quad \sigma \mapsto P_\sigma \quad \text{with } P_\sigma(i,j) = \mathbb{I}\{\sigma(i) = j\}$$

  - Kemeny embedding [Jiao et al., 2016]

$$\mathfrak{S}_n \to \mathbb{R}^{n(n-1)/2}, \quad \sigma \mapsto \phi_\sigma \quad \text{with } \phi_\sigma = \begin{pmatrix} \vdots \\ sign(\sigma(i) - \sigma(j)) \\ \vdots \end{pmatrix}_{i<j}$$

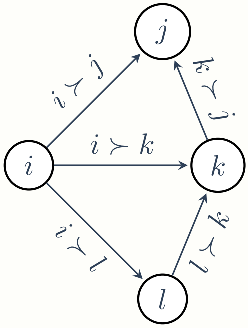
- Harmonic analysis
  - Fourier analysis [Clémençon et al., 2011], [Kondor and Barbosa, 2010]

$$\hat{h}_\lambda = \sum_{\sigma\in\mathfrak{S}_n} h(\sigma)\rho_\lambda(\sigma) \text{ où } \rho_\lambda(\sigma) \in \mathbb{C}^{d_\lambda\times d_\lambda} \text{ for all } \lambda \vdash n.$$

  - Multiresolution analysis for incomplete rankings [Sibony et al., 2015]

Nonparametric approaches - Examples 2

Modeling of pairwise comparisons as a graph:



- HodgeRank exploits the topology of the graph [Jiang et al., 2011]
- Approximation of pairwise comparison matrices [Shah and Wainwright, 2015]

Some ranking problems

Perform some task on a dataset of N rankings $\mathcal{D}_N = (\prec_1, \ldots, \prec_N)$.

Examples

- **Top-1 recovery:** Find the "most preferred" item in $\mathcal{D}_N$
  e.g. Output of an election
- **Aggregation:** Find a full ranking that "best summarizes" $\mathcal{D}_N$
  e.g. Ranking of a competition
- **Clustering:** Split $\mathcal{D}_N$ into clusters
  e.g. Segment customers based on their answers to a survey
- **Prediction:** Predict the outcome of a missing pairwise comparison in a ranking $\prec$
  e.g. In a recommendation setting

Outline

Introduction to Ranking Data

Background on Ranking Aggregation

Ranking Median Regression

Local Consensus Methods for Ranking Median Regression

Conclusion

The Ranking Aggregation Problem

Framework

- n items: $\{1, \dots, n\}$.
- N rankings/permutations : $\Sigma_1, \dots, \Sigma_N$.

Consensus Ranking

Suppose we have a dataset of rankings/permutations $\mathcal{D}_N = (\Sigma_1, \dots, \Sigma_N) \in \mathfrak{S}_n^N$. We want to find a global order ("consensus") σ^* on the n items that best represents the dataset.

Main methods (Non parametric)

- Scoring methods: Copeland, Borda
- Metric-based method: Kemeny's rule

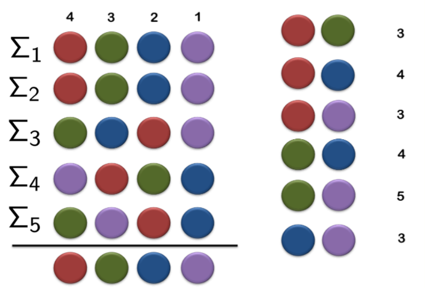
Methods for Ranking Aggregation

Copeland method

Sort the items according to their Copeland score, defined for each item i by:

$$s_C(i) = \frac{1}{N} \sum_{t=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{n} \mathbb{I}[\Sigma_t(i) < \Sigma_t(j)]$$

which counts the number of pairwise victories of item i over the other items $j \neq i$.



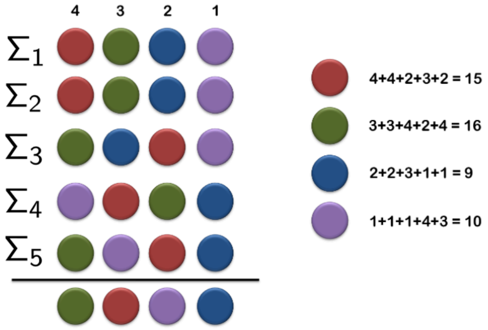
Methods for Ranking Aggregation

Borda Count

Sort the items according to their Borda score, defined for each item i by:

$$s_B(i) = \frac{1}{N}\sum_{t=1}^{N}(n+1-\Sigma_t(i))$$

which is "the average" rank of item i.



Methods for Ranking Aggregation

Kemeny's rule (1959)

Find the solution of :

$$\min_{\sigma \in \mathfrak{S}_n} \sum_{t=1}^{N} d(\sigma, \Sigma_t)$$

where d is the Kendall's tau distance:

$$d_\tau(\sigma, \Sigma) = \sum_{i<j} \mathbb{I}\{(\sigma(i) - \sigma(j))(\Sigma(i) - \Sigma(j)) < 0\},$$

which counts the number of pairwise disagreements (or minimal number of adjacent transpositions to convert σ into Σ).

Ex: σ= 1234, Σ= 2413 $\Rightarrow d_\tau(\sigma, \Sigma) = 3$ (disagree on 12,14,34).

Kemeny's rule

Kemeny's consensus has a lot of interesting properties:

- Social choice justification: Satisfies many voting properties, such as the **Condorcet criterion**: if an alternative is preferred to all others in pairwise comparisons then it is the winner [Young and Levenglick, 1978]
- Statistical justification: Outputs the maximum likelihood estimator under the Mallows model [Young, 1988]
- Main drawback: NP-hard in the number of items n [Bartholdi et al., 1989] even for $N = 4$ votes [Dwork et al., 2001].

Our contribution: we give conditions for the exact Kemeny aggregation to become tractable [Korba et al., 2017].

Statistical Ranking Aggregation

Kemeny's rule:

$$\min_{\sigma \in \mathfrak{S}_n} \sum_{t=1}^{N} d(\sigma, \Sigma_t) \tag{1}$$

Probabilistic Modeling:

$$\mathcal{D}_N = (\Sigma_1, \ldots, \Sigma_N) \quad \text{with} \quad \Sigma_t \sim P$$

Definition

A **Kemeny median** of P is solution of:

$$\min_{\sigma \in \mathfrak{S}_n} L_P(\sigma),$$

where $L_P(\sigma) = \mathbb{E}_{\Sigma \sim P}[d(\Sigma, \sigma)]$ is **the risk** of σ.

Notations:

Let $\sigma^*_P = \text{argmin}_{\sigma \in \mathfrak{S}_n} L_P(\sigma)$ and $\sigma^*_{\widehat{P}_N} = \text{argmin}_{\sigma \in \mathfrak{S}_n} L_{\widehat{P}_N}(\sigma)$ (1)
where $\widehat{P}_N = \frac{1}{N} \sum_{k=1}^{N} \delta_{\Sigma_i}$.

Risk of Ranking Aggregation

The risk of a median σ is $L(\sigma) = \mathbb{E}_{\Sigma \sim P}[d(\Sigma, \sigma)]$, where d is:

$$d(\sigma, \sigma') = \sum_{\{i,j\} \subset [\![n]\!]} \{(\sigma(i) - \sigma(j))(\sigma'(i) - \sigma'(j)) < 0\}$$

Let $p_{i,j} = \mathbb{P}[\Sigma(i) < \Sigma(j)]$ the probability that item i is preferred to item j.

Risk of Ranking Aggregation

The risk of a median σ is $L(\sigma) = \mathbb{E}_{\Sigma \sim P}[d(\Sigma, \sigma)]$, where d is:

$$d(\sigma, \sigma') = \sum_{\{i,j\} \subset [\![n]\!]} \{(\sigma(i) - \sigma(j))(\sigma'(i) - \sigma'(j)) < 0\}$$

Let $p_{i,j} = \mathbb{P}[\Sigma(i) < \Sigma(j)]$ the probability that item i is preferred to item j.

The risk can be rewritten:

$$L(\sigma) = \sum_{i<j} p_{i,j} \mathbb{I}\{\sigma(i) > \sigma(j)\} + \sum_{i<j} (1 - p_{i,j}) \mathbb{I}\{\sigma(i) < \sigma(j)\}.$$

So if there exists a permutation σ verifying: $\forall i < j$ s.t. $p_{i,j} \neq 1/2$,

$$(\sigma(j) - \sigma(i)) \cdot (p_{i,j} - 1/2) > 0,$$

it would be necessary a median $\sigma^*_P = \text{argmin}_{\sigma \in \mathfrak{S}_n} L_P(\sigma)$ for P.

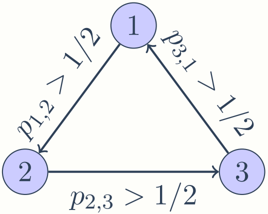
Conditions for Optimality

- the Stochastic Transitivity condition:

$$p_{i,j} \geq 1/2 \text{ and } p_{j,k} \geq 1/2 \Rightarrow p_{i,k} \geq 1/2.$$

  In addition, if $p_{i,j} \neq 1/2$ for all $i < j$, P is said to be "strictly stochastically transitive"" (**SST**)

  $\Rightarrow$ prevents **cycles**:



  $\Rightarrow$ includes Plackett-Luce, Mallows...

- the Low-Noise condition **NA**(h) for some $h > 0$:

$$\min_{i<j} |p_{i,j} - 1/2| \geq h.$$

Main Results [Korba et al., 2017]

- **Optimality.** If P satisfies **SST**, its Kemeny median is **unique** and is given by its Copeland ranking:

$$\sigma_P^*(i) = 1 + \sum_{j \neq i} \mathbb{I}\{p_{i,j} < \frac{1}{2}\}$$

Main Results [Korba et al., 2017]

- **Optimality.** If P satisfies **SST**, its Kemeny median is **unique** and is given by its Copeland ranking:

$$\sigma^*_P(i) = 1 + \sum_{j \neq i} \mathbb{I}\{p_{i,j} < \frac{1}{2}\}$$

- **Generalization.** Then, if P satisfies **SST and NA**(h) for a given $h > 0$, the empirical Copeland ranking:

$$\widehat{s}_N(i) = 1 + \sum_{j \neq i} \mathbb{I}\{\widehat{p}_{i,j} < \frac{1}{2}\} \quad \text{for } 1 \leq i \leq n$$

is in $\mathfrak{S}_n$ and $\widehat{s}_N = \sigma^*_{\widehat{P}_N} = \sigma^*_P$ with overwhelming probability $1 - \frac{n(n-1)}{4} e^{-\alpha_h N}$ with $\alpha_h = \frac{1}{2} \log\left(1/(1-4h^2)\right)$.

$\Rightarrow$ Under the needed conditions, empirical Copeland method ($\mathcal{O}(N\binom{n}{2})$) outputs the true Kemeny consensus (NP-hard) with high probability!

Outline

Introduction to Ranking Data

Background on Ranking Aggregation

Ranking Median Regression

Local Consensus Methods for Ranking Median Regression

Conclusion

Our Problem

Suppose we observe $(X_1, \Sigma_1), \ldots, (X_N, \Sigma_N)$ i.i.d. copies of the pair $(X,\ \Sigma)$, where

- $X \sim \mu$, where μ is a distribution on some feature space $\mathcal{X}$
- $\Sigma \sim P_X$, where P_X is the conditional probability distribution (on $\mathfrak{S}_n$): $P_X(\sigma) = \mathbb{P}[\Sigma = \sigma | X]$

Ex: Users i with characteristics X_i order items by preference resulting in Σ_i.

Goal: Learn a predictive ranking rule :

$$\begin{array}{cccc} s & : & \mathcal{X} & \rightarrow & \mathfrak{S}_n \\ & & x & \mapsto & s(x) \end{array}$$

which given a random vector X, predicts the permutation Σ on the n items.

Performance: Measured by the risk:

$$\mathcal{R}(s) = \mathbb{E}_{X \sim \mu, \Sigma \sim P_X} \left[d_\tau \left(s(X), \Sigma \right) \right]$$

Related Work

- Has been referred to as **label ranking** in the literature [Tsoumakas et al., 2009], [Vembu and Gärtner, 2010]
  $\rightarrow$ Related to multiclass and multilabel classification
  $\rightarrow$ A lot of applications (bioinformatics, meta-learning...)
- A lot of approaches rely on parametric modelling [Cheng and Hüllermeier, 2009], [Cheng et al., 2009], [Cheng et al., 2010]
- MLE or Bayesian Techniques [Rendle et al., 2009],[Lu and Negahban, 2015]

$\Rightarrow$ We develop an approach free of any parametric assumptions.

Ranking Median Regression Approach

$$\mathcal{R}(s) = \mathbb{E}_{X\sim\mu}\left[\mathbb{E}_{\Sigma\sim P_X}\left[d_\tau\left(s(X),\Sigma\right)\right]\right] = \mathbb{E}_{X\sim\mu}\left[L_{P_X}(s(X))\right] \quad (2)$$

Assumption

For $X \in \mathcal{X}$, P_X is **SST**: $\Rightarrow \sigma^*_{P_X} = \text{argmin}_{\sigma\in\mathfrak{S}_n} L_{P_X}(\sigma)$ is **unique**.

Optimal elements

The predictors s minimizing (2) are the ones that maps any point $X \in \mathcal{X}$ to any **conditional** Kemeny median of P_X:

$$s^* = \underset{s\in\mathcal{S}}{\text{argmin}}\, \mathcal{R}(s) \quad \Leftrightarrow \quad s^*(X) = \sigma^*_{P_X}$$

Ranking Median Regression Approach

$$\mathcal{R}(s) = \mathbb{E}_{X\sim\mu}\left[\mathbb{E}_{\Sigma\sim P_X}\left[d_\tau\left(s(X),\Sigma\right)\right]\right] = \mathbb{E}_{X\sim\mu}\left[L_{P_X}(s(X))\right] \quad (2)$$

Assumption

For $X \in \mathcal{X}$, P_X is **SST**: $\Rightarrow \sigma^*_{P_X} = \text{argmin}_{\sigma\in\mathfrak{S}_n} L_{P_X}(\sigma)$ is **unique**.

Optimal elements

The predictors s minimizing (2) are the ones that maps any point $X \in \mathcal{X}$ to any **conditional** Kemeny median of P_X:

$$s^* = \underset{s\in\mathcal{S}}{\text{argmin}}\, \mathcal{R}(s) \quad \Leftrightarrow \quad s^*(X) = \sigma^*_{P_X}$$

Ranking Median Regression

To minimize (2) approximately, instead of computing $\sigma^*_{P_X}$ for any $X = x$, we relax it to Kemeny medians within a cell $\mathcal{C}$ containing x.

$\Rightarrow$ We develop Local consensus methods.

Statistical Framework- ERM

Consider a statistical version of the theoretical risk based on the training data (X_t, Σ_t)'s:

$$\widehat{\mathcal{R}}_N(s) = \frac{1}{N}\sum_{k=1}^{N} d_\tau(s(X_k),\ \Sigma_k)$$

and solutions of the optimization problem:

$$\min_{s\in\mathcal{S}} \widehat{\mathcal{R}}_N(s),$$

where $\mathcal{S}$ is the set of measurable mappings.

Statistical Framework- ERM

Consider a statistical version of the theoretical risk based on the training data (X_t, Σ_t)'s:

$$\widehat{\mathcal{R}}_N(s) = \frac{1}{N}\sum_{k=1}^{N} d_\tau(s(X_k),\ \Sigma_k)$$

and solutions of the optimization problem:

$$\min_{s\in\mathcal{S}} \widehat{\mathcal{R}}_N(s),$$

where $\mathcal{S}$ is the set of measurable mappings.

$\Rightarrow$ We will consider a subset $\mathcal{S}_{\mathcal{P}} \subset \mathcal{S}$:

- supposed to be rich enough to contain approximate versions of $s^* = \text{argmin}_{s\in\mathcal{S}} \mathcal{R}(s)$ (*i.e.* so that $\inf_{s\in\mathcal{S}_{\mathcal{P}}} \mathcal{R}(s) - \mathcal{R}(s^*)$ is 'small')
- ideally appropriate for continuous or greedy optimization.

Outline

Introduction to Ranking Data

Background on Ranking Aggregation

Ranking Median Regression

Local Consensus Methods for Ranking Median Regression

Conclusion

Piecewise Constant Ranking Rules

Let $\mathcal{P} = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ be a partition of the feature space $\mathcal{X}$.
Let $\mathcal{S}_{\mathcal{P}}$ be the collection of all ranking rules that are constant on each cell of $\mathcal{P}$. Any $s \in \mathcal{S}_{\mathcal{P}}$ can be written as:

$$s_{\mathcal{P},\bar{\sigma}}(x) = \sum_{k=1}^{K} \sigma_k \cdot \mathbb{I}\{x \in \mathcal{C}_k\} \text{ where } \bar{\sigma} = (\sigma_1, \ldots, \sigma_K)$$

Piecewise Constant Ranking Rules

Let $\mathcal{P} = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ be a partition of the feature space $\mathcal{X}$.
Let $\mathcal{S}_\mathcal{P}$ be the collection of all ranking rules that are constant on each cell of $\mathcal{P}$. Any $s \in \mathcal{S}_\mathcal{P}$ can be written as:

$$s_{\mathcal{P},\bar{\sigma}}(x) = \sum_{k=1}^{K} \sigma_k \cdot \mathbb{I}\{x \in \mathcal{C}_k\} \text{ where } \bar{\sigma} = (\sigma_1, \ldots, \sigma_K)$$

Local Learning

Let $P_\mathcal{C}$ the cond. distr. of Σ given $X \in \mathcal{C}$: $P_\mathcal{C}(\sigma) = \mathbb{P}[\Sigma = \sigma | X \in \mathcal{C}]$
Recall: P_X is SST for any $X \in \mathcal{X}$.
Idea: $P_\mathcal{C}$ is still SST and $\sigma^*_{P_\mathcal{C}} = \sigma^*_{P_X}$ provided the $\mathcal{C}_k$'s are small enough.

Theoretical guarantees: Suppose $\exists M < \infty$ s.t. $\forall (x, x') \in \mathcal{X}^2$, $\sum_{i<j} |p_{i,j}(x) - p_{i,j}(x')| \leq \cdot ||x - x'||$, then:

$$\mathcal{R}(s_\mathcal{P}) - \mathcal{R}^* \leq M.\delta_\mathcal{P}$$

where $\delta_\mathcal{P}$ is the max. diameter of $\mathcal{P}$'s cells.

Partitioning Methods

Goal: Generate partitions $\mathcal{P}_N$ in a data-driven fashion.
Two methods tailored to ranking regression are investigated:

- k-nearest neighbor (Voronoi partitioning)
- decision tree (Recursive partitioning)

Local Kemeny Medians

In practice, for a cell $\mathcal{C}$ in $\mathcal{P}_N$, consider $\widehat{P}_{\mathcal{C}} = \frac{1}{N_{\mathcal{C}}} \sum_{k: X_k \in \mathcal{C}} \delta_{\Sigma_k}$, where $N_{\mathcal{C}} = \sum_{k=1}^{N} \mathbb{I}\{X_k \in \mathcal{C}\}$

- If $\widehat{P}_{\mathcal{C}}$ is SST, compute $\sigma^*_{\widehat{P}_{\mathcal{C}}}$ with Copeland method based on $\widehat{p}_{i,j}(\mathcal{C})$
- Else, compute $\widetilde{\sigma}^*_{\widehat{P}_{\mathcal{C}}}$ with empirical Borda count (breaking ties arbitrarily if any)

K-Nearest Neigbors Algorithm

1. Fix $k \in \{1, \ldots, N\}$ and a query point $x \in \mathcal{X}$
2. Sort the training data $(X_1, \Sigma_1), \ldots, (X_N, \Sigma_N)$ by increasing order of the distance to x : $\|X_{(1,N)} - x\| \leq \ldots \leq \|X_{(N,N)} - x\|$
3. Consider next the empirical distribution calculated using the k training points closest to x

$$\widehat{P}(x) = \frac{1}{k} \sum_{l=1}^{k} \delta_{\Sigma_{(l,N)}}$$

and compute the pseudo-empirical Kemeny median, yielding the k-NN prediction at x:

$$s_{k,N}(x) \overset{def}{=} \widetilde{\sigma}^*_{\widehat{P}(x)}.$$

$\Rightarrow$ We recover the classical bound $\mathcal{R}(s_{k,N}) - \mathcal{R}^* = \mathcal{O}(\frac{1}{\sqrt{k}} + \frac{k}{N})$

Decision Tree

Split recursively the feature space $\mathcal{X}$ to minimize some impurity criterion. In each final cell, compute the terminal value based on the data in the cell. Here, for a cell $\mathcal{C} \in \mathcal{P}_N$:

- Impurity:

$$\gamma_{\widehat{P}_{\mathcal{C}}} = \frac{1}{2} \sum_{i<j} \widehat{p}_{i,j}(\mathcal{C}) \left(1 - \widehat{p}_{i,j}(\mathcal{C})\right)$$

  which is tractable and satisfies the double inequality

$$\widehat{\gamma}_{\widehat{P}_{\mathcal{C}}} \leq \min_{\sigma \in \mathfrak{S}_n} L_{\widehat{P}_{\mathcal{C}}}(\sigma) \leq 2\widehat{\gamma}_{\widehat{P}_{\mathcal{C}}}.$$

  Analog to Gini criterion in classification: m classes, f_i proportion of class $i \rightarrow I_G(f) = \sum_{i=1}^{m} f_i(1 - f_i)$

- Terminal value : Compute the pseudo-empirical median of a cell $\mathcal{C}$:

$$s_{\mathcal{C}}(x) \stackrel{def}{=} \widetilde{\sigma}^*_{\widehat{P}_{\mathcal{C}}}.$$

Simulated Data

- We generate two explanatory variables, varying their nature (numerical, categorical) $\Rightarrow$ Setting 1/2/3
- We generate a partition of the feature space
- On each cell of the partition, a dataset of full rankings is generated, varying the distribution (constant, Mallows with $\neq$ dispersion): $D_0/D_1/D_2$

D_i	Setting 1			Setting 2			Setting 3		
	n=3	n=5	n=8	n=3	n=5	n=8	n=3	n=5	n=8
D_0	0.0698* 0.0473** (0.578)	0.1290* 0.136** (1.147)	0.2670* 0.324** (2.347)	0.0173* 0.0568** (0.596)	0.0405* 0.145** (1.475)	0.110* 0.2695** (3.223)	0.0112* 0.099** (0.5012)	0.0372* 0.1331** (1.104)	0.0862* 0.2188** (2.332)
D_1	0.3475 * 0.307** (0.719)	0.569* 0.529** (1.349)	0.9405 * 0.921** (2.606)	0.306* 0.308** (0.727)	0.494* 0.536** (1.634)	0.784* 0.862** (3.424)	0.289* 0.3374** (0.5254)	0.457* 0.5714** (1.138)	0.668* 0.8544** (2.287)
D_2	0.8656* 0.7228** (0.981)	1.522* 1.322** (1.865)	2.503* 2.226** (3.443)	0.8305 * 0.723** (1.014)	1.447 * 1.3305** (2.0945)	2.359* 2.163** (4.086)	0.8105* 0.7312** (0.8504)	1.437* 1.3237** (1.709)	2.189* 2.252** (3.005)

Table: Empirical risk averaged on 50 trials on simulated data.

(): Clustering +PL, *: K-NN, **: Decision Tree

US General Social Survey

Participants were asked to rank 5 aspects about a job: ”high income”, ”no danger of being fired”, ”short working hours”, ”chances for advancement”, ”work important and gives a feeling of accomplishment”.

- 18544 samples collected between 1973 and 2014.
- 8 individual attributes are considered: sex, race, birth cohort, highest educational degree attained, family income, marital status, number of children, household size
- plus 3 attributes of work conditions: working status, employment status, and occupation.

Results:
Risk of decision tree: 2,763 $\rightarrow$ Splitting variables:

1) occupation 2) race 3) degree

Outline

Introduction to Ranking Data

Background on Ranking Aggregation

Ranking Median Regression

Local Consensus Methods for Ranking Median Regression

Conclusion

Conclusion

Ranking data is fun!

Its analysis presents great and interesting challenges:

- Most of the maths from euclidean spaces cannot be applied
- But our intuitions still hold
- Based on our results on ranking aggregation, we develop a novel approach to ranking regression/label ranking

Openings: Extension to pairwise comparisons

Big challenges

- How to extend to incomplete rankings (+with ties)?
- How to extend to items with features?

Bartholdi, J. J., Tovey, C. A., and Trick, M. A. (1989).
The computational difficulty of manipulating an election.
Social Choice and Welfare, 6(3):227–241.

Cheng, W., Dembczyński, K., and Hüllermeier, E. (2010).
Label ranking methods based on the Plackett-Luce model.
In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 215–222.

Cheng, W., Hühn, J., and Hüllermeier, E. (2009).
Decision tree and instance-based learning for label ranking.
In *Proceedings of the 26th International Conference on Machine Learning (ICML-09)*, pages 161–168.

Cheng, W. and Hüllermeier, E. (2009).
A new instance-based label ranking approach using the mallows model.
Advances in Neural Networks–ISNN 2009, pages 707–716.

Clémençon, S., Gaudel, R., and Jakubowicz, J. (2011).

On clustering rank data in the fourier domain.
In *ECML*.

Dwork, C., Kumar, R., Naor, M., and Sivakumar, D. (2001).
Rank aggregation methods for the Web.
In *Proceedings of the 10th International WWW conference*, pages 613–622.

Jiang, X., Lim, L.-H., Yao, Y., and Ye, Y. (2011).
Statistical ranking and combinatorial hodge theory.
Mathematical Programming, 127(1):203–244.

Jiao, Y., Korba, A., and Sibony, E. (2016).
Controlling the distance to a kemeny consensus without computing it.
In *Proceeding of ICML 2016*.

Kondor, R. and Barbosa, M. S. (2010).
Ranking with kernels in Fourier space.
In *Proceedings of COLT'10*, pages 451–463.

Korba, A., Clémençon, S., and Sibony, E. (2017).

A learning theory of ranking aggregation.
In *Proceeding of AISTATS 2017*.

Lu, Y. and Negahban, S. N. (2015).
Individualized rank aggregation using nuclear norm regularization.
In *Communication, Control, and Computing (Allerton), 2015 53rd Annual Allerton Conference on*, pages 1473–1479. IEEE.

Luce, R. D. (1959).
Individual Choice Behavior.
Wiley.

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

Plackett, R. L. (1975).
The analysis of permutations.
Applied Statistics, 2(24):193–202.

Plis, S., McCracken, S., Lane, T., and Calhoun, V. (2011).
Directional statistics on permutations.
In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 600–608.

Rendle, S., Freudenthaler, C., Gantner, Z., and Schmidt-Thieme, L. (2009).
Bpr: Bayesian personalized ranking from implicit feedback.
In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 452–461. AUAI Press.

Shah, N. B. and Wainwright, M. J. (2015).
Simple, robust and optimal ranking from pairwise comparisons.
arXiv preprint arXiv:1512.08949.

Sibony, E., Clémençon, S., and Jakubowicz, J. (2015).
MRA-based statistical learning from incomplete rankings.
In *Proceeding of ICML*.

Thurstone, L. L. (1927).
A law of comparative judgment.
Psychological Review, 34(4):273–286.

Tsoumakas, G., Katakis, I., and Vlahavas, I. (2009).
Mining multi-label data.
In *Data mining and knowledge discovery handbook*, pages 667–685. Springer.

Vembu, S. and Gärtner, T. (2010).
Label ranking algorithms: A survey.
In *Preference learning*, pages 45–64. Springer.

Young, H. (1988).
Condorcet's theory of voting.
American Political Science Review, 82(4):1231–1244.

Young, H. P. and Levenglick, A. (1978).
A consistent extension of condorcet's election principle.
SIAM Journal on applied Mathematics, 35(2):285–300.