

Learning from Ranking Data: Theory and Methods

Anna Korba

LTCI, Télécom ParisTech, Université Paris-Saclay

October 25th, 2018



Jury: Jean-Philippe Vert (*Mines ParisTech*), Nicolas Vayatis (*ENS Cachan*),

Florence d'Alché-Buc (*Télécom Paris*)

(Reviewers): Shivani Agarwal (*University of Pennsylvania*), Eyke Hüllermeier (*University of Paderborn*)

(Director): Stephan Cléménçon (*Télécom ParisTech*)

Outline

1. Introduction to Ranking Data

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

2. Ranking Aggregation - A Statistical Framework

$$\min_{\sigma \in \mathfrak{S}_n} L(\sigma), \text{ with } L(\sigma) = \mathbb{E}[d_\tau(\sigma, \Sigma)]$$

3. Label Ranking - Piecewise Constant Rules

$$\min_{s : \mathcal{X} \rightarrow \mathfrak{S}_n} \mathcal{R}(s), \text{ with } \mathcal{R}(s) = \mathbb{E}[d_\tau(s(X), \Sigma)]$$

4. Label Ranking - A Structured Prediction Approach

$$\min_{s : \mathcal{X} \rightarrow \mathfrak{S}_n} \mathcal{R}(s), \text{ with } \mathcal{R}(s) = \mathbb{E}[\Delta(s(X), \Sigma)]$$

5. Conclusion

Outline

Introduction to Ranking Data

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

Ranking Aggregation - A Statistical Framework

$$\min_{\sigma \in \mathfrak{S}_n} L(\sigma), \text{ with } L(\sigma) = \mathbb{E}[d_\tau(\sigma, \Sigma)]$$

Label Ranking - Piecewise Constant Rules

$$\min_{s : \mathcal{X} \rightarrow \mathfrak{S}_n} \mathcal{R}(s), \text{ with } \mathcal{R}(s) = \mathbb{E}[d_\tau(s(X), \Sigma)]$$

Label Ranking - A Structured Prediction Approach

$$\min_{s : \mathcal{X} \rightarrow \mathfrak{S}_n} \mathcal{R}(s), \text{ with } \mathcal{R}(s) = \mathbb{E}[\Delta(s(X), \Sigma)]$$

Conclusion

What is ranking data?

Consider a set of items $\llbracket n \rrbracket := \{1, \dots, n\}$.

A ranking is an **ordered list** (of any size) **of items** of $\llbracket n \rrbracket$

What is ranking data?

Consider a set of items $\llbracket n \rrbracket := \{1, \dots, n\}$.

A ranking is an **ordered list** (of any size) **of items** of $\llbracket n \rrbracket$

Example:

$\llbracket 4 \rrbracket := \{1, 2, 3, 4\} = \begin{matrix} \text{[Actor 1]} \\ \vdots \\ \text{[Actor 2]} \\ \vdots \\ \text{[Actor 3]} \\ \vdots \\ \text{[Actor 4]} \end{matrix}, \quad .$

Ask an actor to rank/order them by preference (\succ):



Many applications involve rankings/comparisons

- ▶ Modelling human preferences (elections, surveys, online implicit feedback)



⇒ easier to rank than to rate

- ▶ Computer systems (search engines, recommendation systems)

- ▶ Others (competitions, biological data...)

Analysis of full rankings

Set of items $\llbracket n \rrbracket := \{1, \dots, n\}$. Ex: $\{1, 2, 3, 4\}$

- ▶ An individual expresses her preferences as a **full** ranking, i.e a strict order \succ over the whole set $\llbracket n \rrbracket$:

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

Other kind of rankings: **Top-k rankings:** $a_1, \dots, a_k \succ \text{the rest}$, **Pairwise comparisons:**

$$a_1 \succ a_2$$

Analysis of full rankings

Set of items $\llbracket n \rrbracket := \{1, \dots, n\}$. Ex: $\{1, 2, 3, 4\}$

- An individual expresses her preferences as a **full ranking**, i.e a strict order \succ over the whole set $\llbracket n \rrbracket$:

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

Other kind of rankings: **Top-k rankings:** $a_1, \dots, a_k \succ \text{the rest}$, **Pairwise comparisons:**

$$a_1 \succ a_2$$

A full ranking can be seen as the permutation σ that maps an item to its rank:

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

$$2 \succ 1 \succ 3 \succ 4 \quad \Leftrightarrow \quad \sigma = 2134 \quad (\sigma(2) = 1, \sigma(1) = 2, \dots)$$

Analysis of full rankings

Set of items $\llbracket n \rrbracket := \{1, \dots, n\}$. Ex: $\{1, 2, 3, 4\}$

- An individual expresses her preferences as a **full ranking**, i.e a strict order \succ over the whole set $\llbracket n \rrbracket$:

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

Other kind of rankings: **Top-k rankings:** $a_1, \dots, a_k \succ \text{the rest}$, **Pairwise comparisons:**

$$a_1 \succ a_2$$

A full ranking can be seen as the permutation σ that maps an item to its rank:

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

$$2 \succ 1 \succ 3 \succ 4 \quad \Leftrightarrow \quad \sigma = 2134 \quad (\sigma(2) = 1, \sigma(1) = 2, \dots)$$

Let \mathfrak{S}_n be set of permutations of $\llbracket n \rrbracket$, the symmetric group.
Ex: $\mathfrak{S}_4 = 1234, 1324, 1423, \dots, 4321$

Common ranking problems

Consider N individuals expressing their preferences on $\llbracket n \rrbracket$:
⇒ results in a dataset of N rankings/permuations

$$\mathcal{D}_N = (\sigma_1, \sigma_2, \dots, \sigma_N) \in \mathfrak{S}_n^N$$

Common ranking problems

Consider N individuals expressing their preferences on $[\![n]\!]$:
⇒ results in a dataset of N rankings/permutations

$$\mathcal{D}_N = (\sigma_1, \sigma_2, \dots, \sigma_N) \in \mathfrak{S}_n^N$$

Considered here

- ▶ **Ranking aggregation:** Find a full ranking that “best summarizes” \mathcal{D}_N
e.g. Ranking of a competition
- ▶ **Label ranking:** Predict a ranking given some information
e.g. In a recommendation setting

Others

- ▶ **Top-1 recovery:** Find the “most preferred” item in \mathcal{D}_N
e.g. Output of an election
- ▶ **Clustering:** Split \mathcal{D}_N into clusters
e.g. Segment customers based on their answers to a survey

Analysis of ranking data: main challenges

How to analyze a dataset of permutations $\mathcal{D}_N = (\sigma_1, \dots, \sigma_N)$?

Analysis of ranking data: main challenges

How to analyze a dataset of permutations $\mathcal{D}_N = (\sigma_1, \dots, \sigma_N)$?

- ▶ A random permutation $\Sigma \in \mathfrak{S}_n$ can be seen as a random vector $(\Sigma(1), \dots, \Sigma(n)) \in \mathbb{R}^n$...

Analysis of ranking data: main challenges

How to analyze a dataset of permutations $\mathcal{D}_N = (\sigma_1, \dots, \sigma_N)$?

- ▶ A random permutation $\Sigma \in \mathfrak{S}_n$ can be seen as a random vector $(\Sigma(1), \dots, \Sigma(n)) \in \mathbb{R}^n$...
but the random variables $\Sigma(1), \dots, \Sigma(n)$ are highly dependent and the sum $\Sigma + \Sigma'$ is not a random permutation!

Analysis of ranking data: main challenges

How to analyze a dataset of permutations $\mathcal{D}_N = (\sigma_1, \dots, \sigma_N)$?

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

Analysis of ranking data: main challenges

How to analyze a dataset of permutations $\mathcal{D}_N = (\sigma_1, \dots, \sigma_N)$?

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

Analysis of ranking data: main challenges

How to analyze a dataset of permutations $\mathcal{D}_N = (\sigma_1, \dots, \sigma_N)$?

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

Analysis of ranking data: main challenges

How to analyze a dataset of permutations $\mathcal{D}_N = (\sigma_1, \dots, \sigma_N)$?

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

Main approaches 1 - Parametric

- ▶ Choose a predefined generative model on the data and analyze the data through that model
([Lu and Boutilier, 2014, Zhao et al., 2016, Szörényi et al., 2015])

Main approaches 1 - Parametric

- ▶ Choose a predefined generative model on the data and analyze the data through that model

([Lu and Boutilier, 2014, Zhao et al., 2016, Szörényi et al., 2015])

- ▶ **Mallows** [Mallows, 1957]

Parameterized by a central ranking $\sigma_0 \in \mathfrak{S}_n$ and a dispersion parameter $\gamma \in \mathbb{R}^+$

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

- ▶ **Plackett-Luce** [Luce, 1959]

Each item i is parameterized by w_i with $w_i \in \mathbb{R}^+$:

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

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

Main approaches 1 - Parametric

- ▶ Choose a predefined generative model on the data and analyze the data through that model
([Lu and Boutilier, 2014, Zhao et al., 2016, Szörényi et al., 2015])

- ▶ **Mallows** [Mallows, 1957]

Parameterized by a central ranking $\sigma_0 \in \mathfrak{S}_n$ and a dispersion parameter $\gamma \in \mathbb{R}^+$

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

- ▶ **Plackett-Luce** [Luce, 1959]

Each item i is parameterized by w_i with $w_i \in \mathbb{R}^+$:

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

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

- ▶ may fail to hold on real data (see for instance [Davidson and Marschak, 1959, Tversky, 1972] on decision making)

Main approaches 2 -“Non Parametric”

- ▶ Choose a structure on \mathfrak{S}_n and analyze the data with respect to that structure
 - ▶ Harmonic analysis ([Kondor and Barbosa, 2010, Cléménçon et al., 2011, Sibony et al., 2015])
 - ▶ Kernel density smoothing [Sun et al., 2012]
 - ▶ Modeling of pairwise comparisons ([Jiang et al., 2011, Rajkumar and Agarwal, 2014, Shah and Wainwright, 2017])
 - ▶ Kernel methods [Jiao and Vert, 2015]...

Presented contributions of the thesis

Two problems considered:

- ▶ Unsupervised learning: statistical framework and results for **ranking aggregation**
- ▶ Supervised learning: two family of methods for **label ranking**

Our approach (Non parametric)

- ▶ Introduce a rigorous statistical framework for each problem and establish theoretical guarantees
- ▶ Exploit the topology of the graph of pairwise preferences
 $2 \succ 1 \succ 3 \succ 4 \implies 2 \succ 1, 2 \succ 3, 2 \succ 4, 1 \succ 3, 1 \succ 4, 3 \succ 4$
- ▶ Exploit the geometry of well-chosen feature maps for rankings
 $\sigma \mapsto \phi(\sigma) \in \mathbb{R}$

Outline

Introduction to Ranking Data

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

Ranking Aggregation - A Statistical Framework

$$\min_{\sigma \in \mathfrak{S}_n} L(\sigma), \text{ with } L(\sigma) = \mathbb{E}[d_\tau(\sigma, \Sigma)]$$

Label Ranking - Piecewise Constant Rules

$$\min_{s : \mathcal{X} \rightarrow \mathfrak{S}_n} \mathcal{R}(s), \text{ with } \mathcal{R}(s) = \mathbb{E}[d_\tau(s(X), \Sigma)]$$

Label Ranking - A Structured Prediction Approach

$$\min_{s : \mathcal{X} \rightarrow \mathfrak{S}_n} \mathcal{R}(s), \text{ with } \mathcal{R}(s) = \mathbb{E}[\Delta(s(X), \Sigma)]$$

Conclusion

Ranking Aggregation

Consider a dataset of rankings/permuations:

$$\mathcal{D}_N = (\sigma_1, \dots, \sigma_N) \in \mathfrak{S}_n^N$$

We want to find a global order (*consensus*) on the n items:

$$\sigma_{\mathcal{D}_N}^* \in \mathfrak{S}_n$$

that **best represents** the dataset.

Ranking Aggregation

Consider a dataset of rankings/permuations:

$$\mathcal{D}_N = (\sigma_1, \dots, \sigma_N) \in \mathfrak{S}_n^N$$

We want to find a global order (*consensus*) on the n items:

$$\sigma_{\mathcal{D}_N}^* \in \mathfrak{S}_n$$

that **best represents** the dataset.

- ▶ problem introduced in the study of elections systems in *social choice*
- ▶ modern applications (e.g. meta-search engines, label ranking)
- ▶ how to compute $\sigma_{\mathcal{D}_N}^*$?

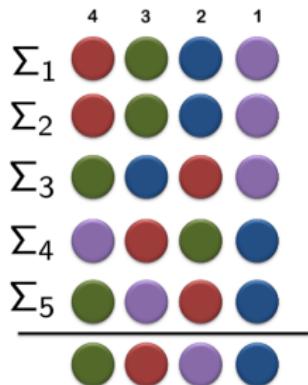
Scoring methods

Idea: compute a score for each item i and sort the items w.r.t. that score.

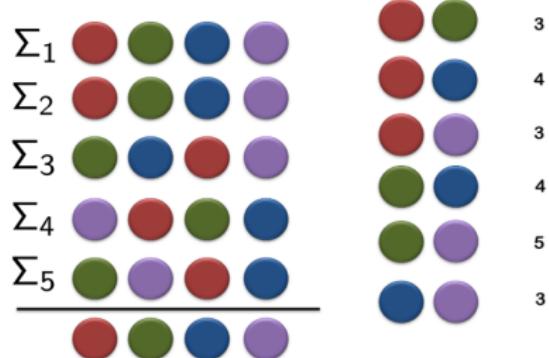
Scoring methods

Idea: compute a score for each item i and sort the items w.r.t. that score.

Borda [Borda, 1781]



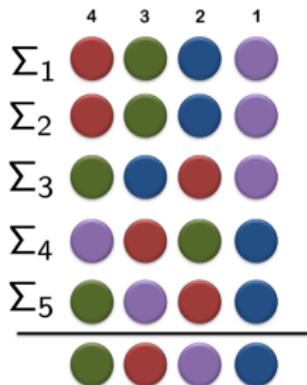
Copeland [Copeland, 1951]



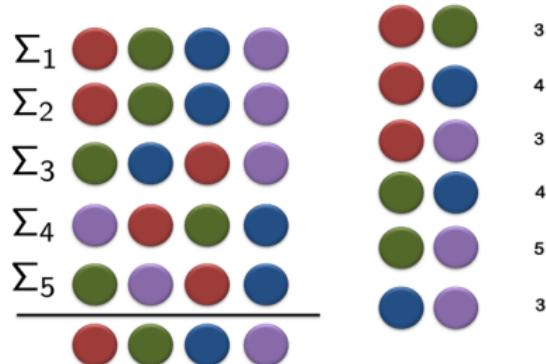
Scoring methods

Idea: compute a score for each item i and sort the items w.r.t. that score.

Borda [Borda, 1781]



Copeland [Copeland, 1951]

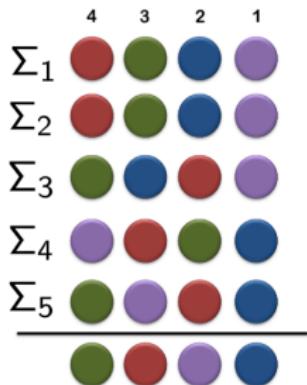


$$s_B(i) = \sum_{k=1}^N (n + 1 - \Sigma_k(i)) = \sum_{\substack{j=1 \\ j \neq i}}^n \sum_{k=1}^N \mathbb{I}[\Sigma_k(i) < \Sigma_k(j)] + N$$

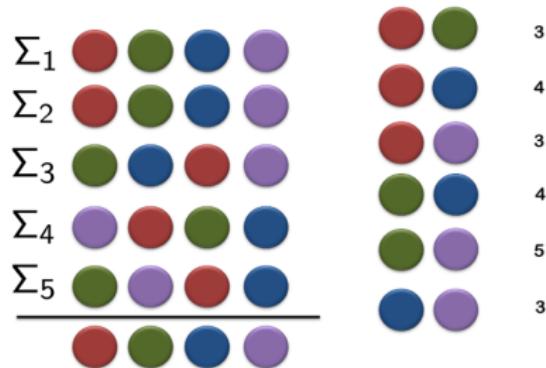
Scoring methods

Idea: compute a score for each item i and sort the items w.r.t. that score.

Borda [Borda, 1781]



Copeland [Copeland, 1951]



$$s_B(i) = \sum_{k=1}^N (n+1 - \Sigma_k(i)) = \sum_{j=1}^n \sum_{\substack{k=1 \\ j \neq i}}^N \mathbb{I}[\Sigma_k(i) < \Sigma_k(j)] + N$$

$$s_C(i) = \sum_{j=1}^n \mathbb{I}\left\{ \sum_{\substack{k=1 \\ j \neq i}}^N \mathbb{I}[\Sigma_k(i) < \Sigma_k(j)] > \frac{N}{2} \right\}$$

Scoring methods

- ▶ Simple complexities (Borda $\mathcal{O}(nN)$, Copeland $\mathcal{O}(n^2N)$)
- ▶ But **few guarantees** on the output
 - ▶ does not necessarily output a full ranking (ties between scores)
 - ▶ **Borda** count does not satisfy the **Condorcet criterion**: if a candidate wins (in majority) against all the others in pairwise duels then it should be ranked first
 - ▶ **Copeland** ignores the relatives strengths of the pairwise comparisons (often many ties)

Scoring methods

- ▶ Simple complexities (Borda $\mathcal{O}(nN)$, Copeland $\mathcal{O}(n^2N)$)
- ▶ But **few guarantees** on the output
 - ▶ does not necessarily output a full ranking (ties between scores)
 - ▶ **Borda** count does not satisfy the **Condorcet criterion**: if a candidate wins (in majority) against all the others in pairwise duels then it should be ranked first
 - ▶ **Copeland** ignores the relatives strengths of the pairwise comparisons (often many ties)

Jean-Charles de Borda Nicolas de Condorcet

⇒ *Borda-Condorcet
debate at the 18th
century*



Kemeny's rule [Kemeny, 1959]

Ranking aggregation as an optimization problem

$$\text{Solve } \sigma_{\mathcal{D}_N}^* = \operatorname*{argmin}_{\sigma \in \mathfrak{S}_n} \sum_{k=1}^N d(\sigma, \Sigma_k)$$

Kemeny's rule [Kemeny, 1959]

Ranking aggregation as an optimization problem

$$\text{Solve } \sigma_{\mathcal{D}_N}^* = \operatorname*{argmin}_{\sigma \in \mathfrak{S}_n} \sum_{k=1}^N d(\sigma, \Sigma_k)$$

where d_τ is the Kendall's τ distance, i.e. for $\sigma, \sigma' \in \mathfrak{S}_n$:

$$d_\tau(\sigma, \sigma') = \sum_{1 \leq i < j \leq n} \mathbb{I}\{(\sigma(i) - \sigma(j))(\sigma'(i) - \sigma'(j)) < 0\},$$

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

Kemeny's rule [Kemeny, 1959]

Ranking aggregation as an optimization problem

$$\text{Solve } \sigma_{\mathcal{D}_N}^* = \operatorname*{argmin}_{\sigma \in \mathfrak{S}_n} \sum_{k=1}^N d(\sigma, \Sigma_k)$$

where d_τ is the Kendall's τ distance, i.e. for $\sigma, \sigma' \in \mathfrak{S}_n$:

$$d_\tau(\sigma, \sigma') = \sum_{1 \leq i < j \leq n} \mathbb{I}\{(\sigma(i) - \sigma(j))(\sigma'(i) - \sigma'(j)) < 0\},$$

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

- ▶ Solution always exists but may not be unique
- ▶ Natural loss d_τ when rankings represent preferences

Kemeny's rule properties

Social choice justification

- ▶ unique rule satisfying:
 - ▶ *Neutrality*: if i and j switch positions in $\Sigma_1, \dots, \Sigma_N$, they should switch positions also in $\sigma_{\mathcal{D}_N}^*$
 - ▶ *Condorcet criterion*: any item which wins every other in pairwise simple majority voting should be ranked first
 - ▶ *Consistency*: if \mathcal{D}_N is split is \mathcal{D}_N^1 and \mathcal{D}_N^2 , and $i \succ j$ in $\sigma_{\mathcal{D}_N^1}^*$ and $\sigma_{\mathcal{D}_N^2}^* \implies i \succ j$ in $\sigma_{\mathcal{D}_N}^*$

Statistical justification

Output the MLE estimator under the Mallows model:

$$\operatorname{argmax}_{\sigma \in \mathfrak{S}_n} \prod_{k=1}^N \frac{e^{-\gamma d(\sigma, \Sigma_k)}}{Z} = \operatorname{argmin}_{\sigma \in \mathfrak{S}_n} \sum_{k=1}^N d_\tau(\Sigma_k, \sigma)$$

Problem: **NP-hard** even for $N = 4$ ([Dwork et al., 2001]).

Statistical Ranking Aggregation [Korba et al., 2017]

Probabilistic Modeling

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

where P distribution on \mathfrak{S}_n .

Statistical Ranking Aggregation [Korba et al., 2017]

Probabilistic Modeling

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

where P distribution on \mathfrak{S}_n .

Definition

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

$$\sigma_P^* = \operatorname{argmin}_{\sigma \in \mathfrak{S}_n} L_P(\sigma), \tag{1}$$

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

⇒ Statistical point of view \neq from social choice: $\sigma_{\mathcal{D}_N}^*$ versus σ_P^*

Statistical Ranking Aggregation [Korba et al., 2017]

Probabilistic Modeling

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

where P distribution on \mathfrak{S}_n .

Definition

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

$$\sigma_P^* = \operatorname{argmin}_{\sigma \in \mathfrak{S}_n} L_P(\sigma), \tag{1}$$

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

⇒ Statistical point of view \neq from social choice: $\sigma_{\mathcal{D}_N}^*$ versus σ_P^*

Question: Can we exhibit some conditions on P so that solving (1) is tractable?

Conditions on the distribution

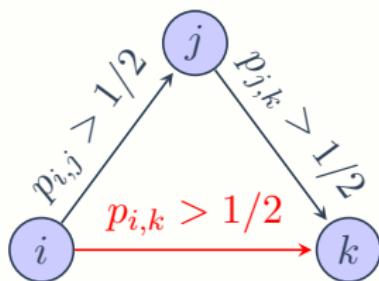
Let $p_{i,j} = \mathbb{P}[\Sigma(i) < \Sigma(j)]$ prob. that item $i \succ j$ (is preferred to).

Conditions on the distribution

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

- Strict Stochastic Transitivity (**SST**): $(p_{i,j} \neq 1/2 \ \forall i, j)$

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

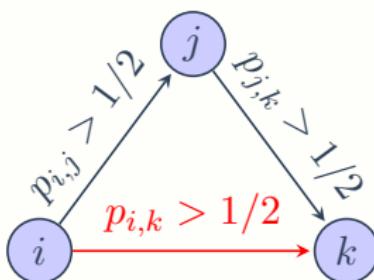


Conditions on the distribution

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

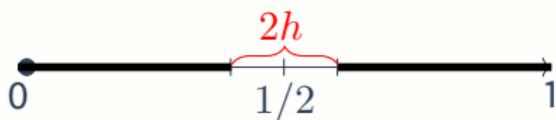
- Strict Stochastic Transitivity (**SST**): $(p_{i,j} \neq 1/2 \forall i, j)$

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



- Low-Noise/**NA**(h) for $h > 0$ ([Audibert and Tsybakov, 2007]):

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



Exact Solutions

Our result [Korba et al., 2017]

Suppose P satisfies **SST** and **NA(h)** for a given $h > 0$. Then with overwhelming probability $1 - \frac{n(n-1)}{4} e^{-\alpha_h N}$:

Exact Solutions

Our result [Korba et al., 2017]

Suppose P satisfies **SST** and **NA(h)** for a given $h > 0$. Then with overwhelming probability $1 - \frac{n(n-1)}{4} e^{-\alpha_h N}$: \hat{P} also verifies **SST**...

Exact Solutions

Our result [Korba et al., 2017]

Suppose P satisfies **SST** and **NA(h)** for a given $h > 0$. Then with overwhelming probability $1 - \frac{n(n-1)}{4}e^{-\alpha_h N}$: \hat{P} also verifies

SST...and the Kemeny median of P is given by the empirical Copeland ranking:

$$\sigma_{\textcolor{red}{P}}^*(i) = 1 + \sum_{j \neq i} \mathbb{I}\{\hat{p}_{i,j} < \frac{1}{2}\} \quad \text{for } 1 \leq i \leq n$$

$$\alpha_h = \frac{1}{2} \log \left(1/(1 - 4h^2) \right)$$

Exact Solutions

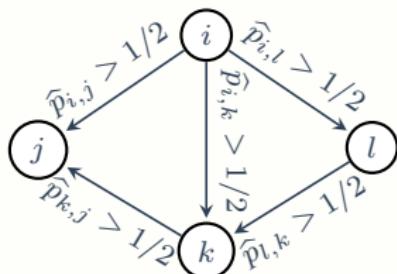
Our result [Korba et al., 2017]

Suppose P satisfies **SST** and **NA(h)** for a given $h > 0$. Then with overwhelming probability $1 - \frac{n(n-1)}{4} e^{-\alpha_h N}$: \hat{P} also verifies

SST...and the Kemeny median of P is given by the empirical Copeland ranking:

$$\sigma_{\hat{P}}^*(i) = 1 + \sum_{j \neq i} \mathbb{I}\{\hat{p}_{i,j} < \frac{1}{2}\} \quad \text{for } 1 \leq i \leq n$$

$$\alpha_h = \frac{1}{2} \log \left(1/(1 - 4h^2) \right)$$



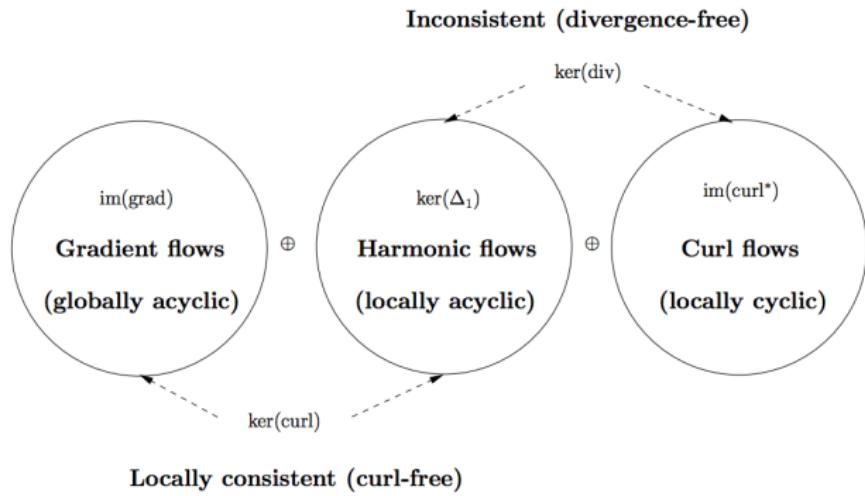
⇒ sort vertices by increasing input degree

Transitivity in practice

- If \hat{P} is SST, compute $\sigma_{\hat{P}}^*$ with Copeland method based on the $\hat{p}_{i,j}$'s

Transitivity in practice

- If \hat{P} is SST, compute $\sigma_{\hat{P}}^*$ with Copeland method based on the $\hat{p}_{i,j}$'s



- Else, compute $\tilde{\sigma}_{\hat{P}}^*$ with empirical Borda count ([Jiang et al., 2011])

$$\tilde{\sigma}_{\hat{P}}^*(i) = \sigma_{proj_{im(grad)}(\hat{P})}^*(i) = \frac{1}{N} \sum_{k=1}^N \Sigma_k(i) \quad \text{for } 1 \leq i \leq n$$

Outline

Introduction to Ranking Data

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

Ranking Aggregation - A Statistical Framework

$$\min_{\sigma \in \mathfrak{S}_n} L(\sigma), \text{ with } L(\sigma) = \mathbb{E}[d_\tau(\sigma, \Sigma)]$$

Label Ranking - Piecewise Constant Rules

$$\min_{s : \mathcal{X} \rightarrow \mathfrak{S}_n} \mathcal{R}(s), \text{ with } \mathcal{R}(s) = \mathbb{E}[d_\tau(s(X), \Sigma)]$$

Label Ranking - A Structured Prediction Approach

$$\min_{s : \mathcal{X} \rightarrow \mathfrak{S}_n} \mathcal{R}(s), \text{ with } \mathcal{R}(s) = \mathbb{E}[\Delta(s(X), \Sigma)]$$

Conclusion

Label Ranking

Now $\mathcal{D}_N = (X_1, \Sigma_1), \dots, (X_N, \Sigma_N)$ i.i.d. copies of (X, Σ)

Label Ranking

Now $\mathcal{D}_N = (X_1, \Sigma_1), \dots, (X_N, \Sigma_N)$ i.i.d. copies of (X, Σ)

Ex: Users i with characteristics X_i and their produced rankings/preferences $\rightarrow \Sigma_i$.

Label Ranking

Now $\mathcal{D}_N = (X_1, \Sigma_1), \dots, (X_N, \Sigma_N)$ i.i.d. copies of (X, Σ)

Ex: Users i with characteristics X_i and their produced rankings/preferences $\rightarrow \Sigma_i$.

- ▶ $X \sim \mu$, 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): $P_X(\sigma) = \mathbb{P}[\Sigma = \sigma | X]$

Label Ranking

Now $\mathcal{D}_N = (X_1, \Sigma_1), \dots, (X_N, \Sigma_N)$ i.i.d. copies of (X, Σ)

Ex: Users i with characteristics X_i and their produced rankings/preferences $\rightarrow \Sigma_i$.

- ▶ $X \sim \mu$, 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): $P_X(\sigma) = \mathbb{P}[\Sigma = \sigma | X]$

Goal: Learn a predictive ranking rule :

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

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



Example: targeted advertising domain

Related Work

- ▶ Can be seen as an extension of multiclass and multilabel classification
- ▶ Many applications, e.g : document categorization, meta-learning
 - ▶ rank a set of topics relevant for a given document
 - ▶ rank a set of algorithms according to their suitability for a new dataset, based on the characteristics of the dataset
- ▶ A lot of approaches rely on parametric modelling
[Cheng and Hüllermeier, 2009], [Cheng et al., 2010]

Related Work

- ▶ Can be seen as an extension of multiclass and multilabel classification
- ▶ Many applications, e.g : document categorization, meta-learning
 - ▶ rank a set of topics relevant for a given document
 - ▶ rank a set of algorithms according to their suitability for a new dataset, based on the characteristics of the dataset
- ▶ A lot of approaches rely on parametric modelling
[Cheng and Hüllermeier, 2009], [Cheng et al., 2010]
⇒ We develop an approach free of any parametric assumptions
(local learning) relying on results and framework developed for
ranking aggregation.

Label Ranking and Ranking Aggregation

Performance: Measured by the risk:

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

where d_{τ} is the Kendall's τ distance, i.e. for $\sigma, \sigma' \in \mathfrak{S}_n$:

$$d_{\tau}(\sigma, \sigma') = \sum_{1 \leq i < j \leq n} \mathbb{I}\{(\sigma(i) - \sigma(j))(\sigma'(i) - \sigma'(j)) < 0\},$$

Label Ranking and Ranking Aggregation

Performance: Measured by the risk:

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

where d_{τ} is the Kendall's τ distance, i.e. for $\sigma, \sigma' \in \mathfrak{S}_n$:

$$d_{\tau}(\sigma, \sigma') = \sum_{1 \leq i < j \leq n} \mathbb{I}\{(\sigma(i) - \sigma(j))(\sigma'(i) - \sigma'(j)) < 0\},$$

Label ranking is an extension of ranking aggregation

$$\begin{aligned} \mathcal{R}(s) &= \mathbb{E}_{X \sim \mu, \Sigma \sim P_X} [d_{\tau}(s(X), \Sigma)] \\ &= \mathbb{E}_{X \sim \mu} [\mathbb{E}_{\Sigma \sim P_{\mathbf{X}}} [d_{\tau}(s(X), \Sigma)]] \\ &= \mathbb{E}_{X \sim \mu} [L_{P_{\mathbf{X}}}(s(X))] \end{aligned}$$

Recall: $\sigma_{P_{\mathbf{X}}}^* = \operatorname{argmin}_{\sigma \in \mathfrak{S}_n} L_{P_{\mathbf{X}}}(\sigma)$

Optimal Elements and Relaxation

Assumption

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

Optimal Elements and Relaxation

Assumption

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

Optimal elements

The best label ranking rules s^* (minimizing $\mathcal{R}(s)$) are the ones that maps any point $X \in \mathcal{X}$ to the **conditional** Kemeny median of P_X :

$$s^* = \operatorname{argmin}_{s: \mathcal{X} \rightarrow \mathfrak{S}_n} \mathcal{R}(s) \Leftrightarrow s^*(X) = \sigma_{P_X}^*$$

Optimal Elements and Relaxation

Assumption

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

Optimal elements

The best label ranking rules s^* (minimizing $\mathcal{R}(s)$) are the ones that maps any point $X \in \mathcal{X}$ to the **conditional** Kemeny median of P_X :

$$s^* = \operatorname{argmin}_{s: \mathcal{X} \rightarrow \mathfrak{S}_n} \mathcal{R}(s) \Leftrightarrow s^*(X) = \sigma_{P_X}^*$$

To minimize the risk $\mathcal{R}(s)$ approximately:

$$\sigma_{P_X}^* \text{ for any } X \xrightarrow{\text{is relaxed to}} s(X) = \sigma_{P_C}^* \text{ for any } X \in \mathcal{C}$$

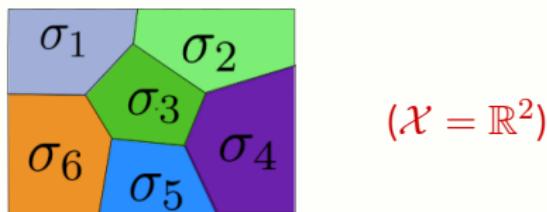
where $P_C(\sigma) = \mathbb{P}[\Sigma = \sigma | X \in \mathcal{C}]$.

\Rightarrow We develop **Piecewise constant rules/ Local consensus methods**.

Piecewise Constant Ranking Rules

Our approach: build a *piecewise constant* ranking rule $s_{\mathcal{P}} \in \mathcal{S}_{\mathcal{P}}$, ie:
constant on each cell of a partition $\mathcal{P} = (\mathcal{C}_1, \dots, \mathcal{C}_K)$ of the feature space \mathcal{X} .

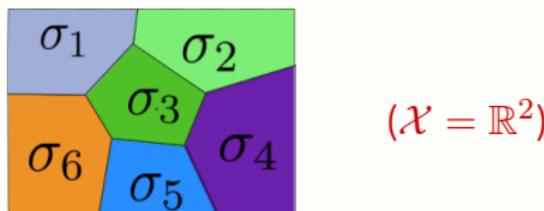
$$\Rightarrow s_{\mathcal{P}}(x) = \sigma_k (\sigma_{P_{\mathcal{C}_k}}^*) \text{ iff } x \in \mathcal{C}_k \text{ for } k = 1, \dots, K.$$



Piecewise Constant Ranking Rules

Our approach: build a *piecewise constant* ranking rule $s_{\mathcal{P}} \in \mathcal{S}_{\mathcal{P}}$, ie: constant on each cell of a partition $\mathcal{P} = (\mathcal{C}_1, \dots, \mathcal{C}_K)$ of the feature space \mathcal{X} .

$$\Rightarrow s_{\mathcal{P}}(x) = \sigma_k (\sigma_{P_{\mathcal{C}_k}}^*) \text{ iff } x \in \mathcal{C}_k \text{ for } k = 1, \dots, K.$$



Approximation Error [Cléménçon et al., 2018]

Suppose that $\exists M < \infty$ such that:

$$\forall (x, x') \in \mathcal{X}^2, \quad \sum_{i < j} |p_{i,j}(x) - p_{i,j}(x')| \leq M \cdot \|x - x'\|.$$

$$\Rightarrow \inf_{s \in \mathcal{S}_{\mathcal{P}}} \mathcal{R}(s) - \mathcal{R}(s^*) \leq M \cdot \delta_{\mathcal{P}}$$

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

Statistical Framework- ERM

Input: training data $\mathcal{D}_N = (X_1, \Sigma_1), \dots, (X_N, \Sigma_N)$

Output: A partition $\mathcal{P}_N = (\mathcal{C}_1, \dots, \mathcal{C}_K)$ of \mathcal{X} built from \mathcal{D}_N .

ERM: Optimize a statistical version of the theoretical risk based on the training data (X_k, Σ_k) 's:

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

Rates of convergence [Cléménçon et al., 2018]

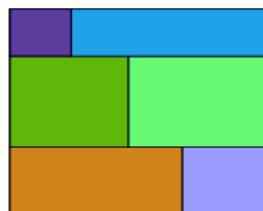
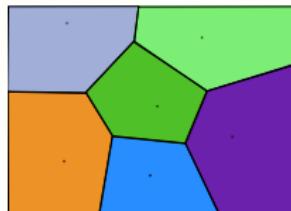
- ▶ classical rates $\mathcal{O}(1/\sqrt{N})$ for ERM.
- ▶ fast rates $\mathcal{O}(1/N)$ under a "uniform" Low-Noise **NA**(h):

$$\inf_{x \in \mathcal{X}} \min_{i < j} |p_{i,j}(x) - 1/2| \geq h.$$

Partitioning Methods

Two methods are investigated:

K-nearest neighbors
(Voronoi partitioning) *Decision tree*
(Recursive partition)

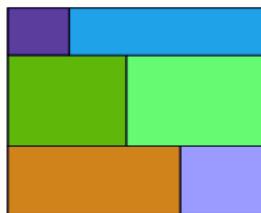
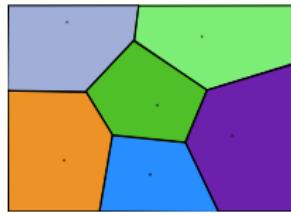


Partitioning Methods

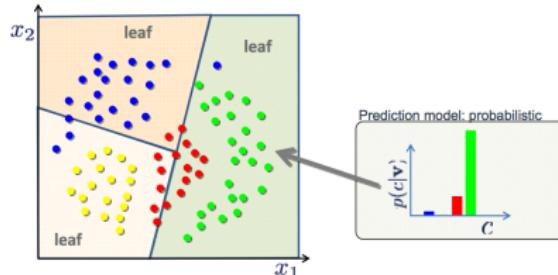
Two methods are investigated:

K-nearest neighbors
(Voronoi partitioning)

Decision tree
(Recursive partition)



For **classification**, the label of a cell (ex: a leaf) is the **majority** label among the training data which fall in this cell.



Compute Local Labels/Medians

Problem: Our training data are *permutations* Σ :

For a cell \mathcal{C} , if $\Sigma_1, \dots, \Sigma_N \in \mathcal{C}$, how do we aggregate them into a final label σ^* ?

⇒ Ranking aggregation problem.

Compute Local Labels/Medians

Problem: Our training data are *permutations* Σ :

For a cell \mathcal{C} , if $\Sigma_1, \dots, \Sigma_N \in \mathcal{C}$, how do we aggregate them into a final label σ^* ?

⇒ Ranking aggregation problem.

For $\mathcal{C} \in \mathcal{P}_N$, consider its empirical distribution:

$$\hat{P}_{\mathcal{C}} = \frac{1}{N_{\mathcal{C}}} \sum_{k: X_k \in \mathcal{C}} \delta_{\Sigma_k}$$

and solve:

$$\min_{\sigma \in \mathfrak{S}_n} L_{\hat{P}_{\mathcal{C}}}(\sigma)$$

⇒ compute locally its Empirical Kemeny median $\tilde{\sigma}_{\hat{P}_{\mathcal{C}}}^*$ with the Copeland method) if $\hat{P}_{\mathcal{C}}$ is SST or Borda otherwise.

K-Nearest Neighbors Algorithm

1. Fix $k \in \{1, \dots, N\}$ and a query point $x \in \mathcal{X}$
2. Sort $(X_1, \Sigma_1), \dots, (X_N, \Sigma_N)$ by increasing order of the distance to x : $\|X_{(1,N)} - x\| \leq \dots \leq \|X_{(N,N)} - x\|$
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-emirical Kemeny median, yielding the k -NN prediction at x :

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

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

Decision Tree

Split recursively the feature space \mathcal{X} by minimizing some impurity criterion.

Decision Tree

Split recursively the feature space \mathcal{X} by minimizing some impurity criterion.

Gini criterion in multiclassification: m classes, f_i proportion of class $i \rightarrow I_G(\mathcal{C}) = \sum_{i=1}^m f_i(\mathcal{C})(1 - f_i(\mathcal{C}))$

Decision Tree

Split recursively the feature space \mathcal{X} by minimizing some impurity criterion.

Gini criterion in multiclassification: m classes, f_i proportion of class

$$i \rightarrow I_G(\mathcal{C}) = \sum_{i=1}^m f_i(\mathcal{C})(1 - f_i(\mathcal{C}))$$

Here, for a cell $\mathcal{C} \in \mathcal{P}_N$:

- ▶ Impurity [Alvo and Yu, 2014]:

$$\gamma_{\hat{P}_{\mathcal{C}}} = \frac{1}{2} \sum_{1 \leq i < j \leq n} \hat{p}_{i,j}(\mathcal{C}) (1 - \hat{p}_{i,j}(\mathcal{C}))$$

(ordering n elements can be seen as $\binom{n}{2}$ classification tasks)
which is tractable and satisfies the double inequality

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

- ▶ Label of a cell/leaf : Compute the pseudo-emirical median of a cell \mathcal{C} :

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

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*	0.1290*	0.2670*	0.0173*	0.0405*	0.110*	0.0112*	0.0372*	0.0862*
	0.0473** (0.578)	0.136** (1.147)	0.324** (2.347)	0.0568** (0.596)	0.145** (1.475)	0.2695** (3.223)	0.099** (0.5012)	0.1331** (1.104)	0.2188** (2.332)
	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

Outline

Introduction to Ranking Data

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

Ranking Aggregation - A Statistical Framework

$$\min_{\sigma \in \mathfrak{S}_n} L(\sigma), \text{ with } L(\sigma) = \mathbb{E}[d_\tau(\sigma, \Sigma)]$$

Label Ranking - Piecewise Constant Rules

$$\min_{s : \mathcal{X} \rightarrow \mathfrak{S}_n} \mathcal{R}(s), \text{ with } \mathcal{R}(s) = \mathbb{E}[d_\tau(s(X), \Sigma)]$$

Label Ranking - A Structured Prediction Approach

$$\min_{s : \mathcal{X} \rightarrow \mathfrak{S}_n} \mathcal{R}(s), \text{ with } \mathcal{R}(s) = \mathbb{E}[\Delta(s(X), \Sigma)]$$

Conclusion

Structured prediction for Label Ranking

Goal: Learn a predictive ranking rule $s : \mathcal{X} \rightarrow \mathfrak{S}_n$ as:

$$\min_{s : \mathcal{X} \rightarrow \mathfrak{S}_n} \mathcal{R}(s), \text{ with } \mathcal{R}(s) = \mathbb{E} [\Delta(s(X), \Sigma)]$$

with Δ some loss function, e.g.:

- ▶ Kendall's τ : $\Delta_\tau(\sigma, \sigma') = \sum_{i < j} \mathbb{I}[(\sigma(i) - \sigma(j))(\sigma'(i) - \sigma'(j)) < 0]$
- ▶ Hamming: $\Delta_H(\sigma, \sigma') = \sum_{i=1}^n \mathbb{I}[\sigma(i) \neq \sigma'(i)].$
Popular when rankings represents matchings/assignments

Structured prediction for Label Ranking

Goal: Learn a predictive ranking rule $s : \mathcal{X} \rightarrow \mathfrak{S}_n$ as:

$$\min_{s : \mathcal{X} \rightarrow \mathfrak{S}_n} \mathcal{R}(s), \text{ with } \mathcal{R}(s) = \mathbb{E} [\Delta(s(X), \Sigma)]$$

with Δ some loss function, e.g.:

- ▶ Kendall's τ : $\Delta_\tau(\sigma, \sigma') = \sum_{i < j} \mathbb{I}[(\sigma(i) - \sigma(j))(\sigma'(i) - \sigma'(j)) < 0]$
- ▶ Hamming: $\Delta_H(\sigma, \sigma') = \sum_{i=1}^n \mathbb{I}[\sigma(i) \neq \sigma'(i)].$
Popular when rankings represents matchings/assignments

Idea [Korba et al., 2018]: Consider a family of Δ loss functions:

$$\Delta(\sigma, \sigma') = \|\phi(\sigma) - \phi(\sigma')\|_{\mathcal{F}}^2. \quad (2)$$

with $\phi : \mathfrak{S}_n \rightarrow \mathcal{F}$ some ranking embedding, i.e. that maps the permutations $\sigma \in \mathfrak{S}_n$ into a Hilbert space \mathcal{F} (e.g. \mathbb{R}^d for $d \in \mathbb{N}$).

Motivation: There exist ϕ_τ, ϕ_H such that Δ_τ and Δ_H write as (2).

Structured Prediction Approach

$$\min_{\mathbf{s} : \mathcal{X} \rightarrow \mathfrak{S}_n} \mathcal{R}(s), \text{ with } \mathcal{R}(s) = \mathbb{E} [\|\phi(\mathbf{s}(X)) - \phi(\Sigma)\|_{\mathcal{F}}^2]$$

⇒ Hard to optimize.

Idea: Introduce a surrogate problem:

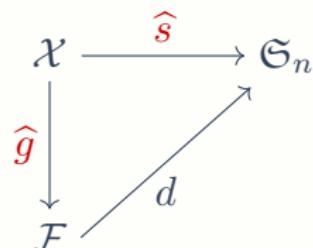
$$\min_{\mathbf{g} : \mathcal{X} \rightarrow \mathcal{F}} \mathcal{L}(g), \quad \text{with} \quad \mathcal{L}(g) = \mathbb{E} [\|\mathbf{g}(X) - \phi(\Sigma)\|_{\mathcal{F}}^2]$$

⇒ easier to optimize since \mathbf{g} has values in \mathcal{F}

Structured Prediction Approach

Firstly pick a loss Δ (or embedding ϕ)

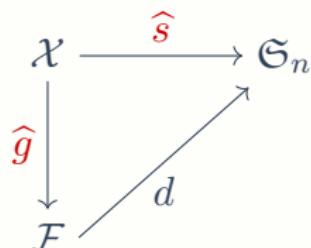
We can thus approach structured prediction in **two steps**:
(see [Ciliberto et al., 2016, Brouard et al., 2016])



Structured Prediction Approach

Firstly pick a loss Δ (or embedding ϕ)

We can thus approach structured prediction in **two steps**:
(see [Ciliberto et al., 2016, Brouard et al., 2016])



- ▶ **Step 1 (Regression):** Learn g from $\mathcal{D}_N = (X_1, \Sigma_1), \dots, (X_N, \Sigma_N)$ with any regression method (kNN, RF, Ridge regression...)
⇒ Output $\hat{g} : \mathcal{X} \rightarrow \mathcal{F}$
- ▶ **Step 2 (Pre-image):** for any $x \in \mathcal{X}$:

$$\hat{s}(x) = d \circ \hat{g}(x) = \operatorname{argmin}_{\sigma \in \mathfrak{S}_n} \|\phi(\sigma) - \hat{g}(x)\|_{\mathcal{F}}^2$$

⇒ Choice of ϕ and regression method matter

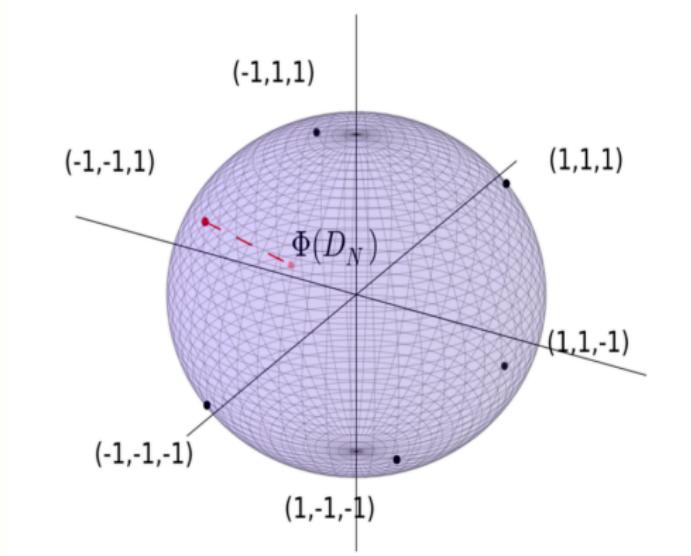
Ranking embeddings proposed - 1

- Kemeny embedding ([Jiao and Vert, 2015, Jiao et al., 2016])

$$\phi_\tau: \mathfrak{S}_n \rightarrow \mathbb{R}^{n(n-1)/2}$$

$$\sigma \mapsto (\text{sign}(\sigma(j) - \sigma(i)))_{1 \leq i < j \leq n} .$$

Ex: $\sigma = 132 \Rightarrow \phi_\tau(\sigma) = (1, 1, -1)$



Ranking embeddings proposed - 2

- ▶ Hamming embedding ([Plis et al., 2011])

$$\begin{aligned}\phi_H: \mathfrak{S}_n &\rightarrow \mathbb{R}^{n \times n} \\ \sigma &\mapsto (\mathbb{I}\{\sigma(i) = j\})_{1 \leq i,j \leq n},\end{aligned}$$

Ex: $\sigma = 132 \Rightarrow \phi_H(\sigma) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$

- ▶ Lehmer embedding ([Li et al., 2017])

$$\begin{aligned}\phi_L: \mathfrak{S}_n &\rightarrow \mathbb{R}^n \\ \sigma &\mapsto (\#\{i : i < j, \sigma(i) > \sigma(j)\})_{j=1,\dots,n},\end{aligned}$$

Ex: $\sigma = 132 \Rightarrow \phi_L(\sigma) = (0, 1, 0)$

Theoretical and computational advantages

For Kemeny and Hamming embedding:

- ▶ **consistency holds:**

$$\mathcal{R}(d \circ \hat{g}) - \mathcal{R}(s^*) \leq c_\phi \sqrt{\mathcal{L}(\hat{g}) - \mathcal{L}(g^*)}$$

with $c_{\phi_\tau} = \sqrt{\frac{n(n-1)}{2}}$ and $c_{\phi_H} = \sqrt{n}$ (constants with n)

- ▶ but the **pre-image step is hard** : NP-hard for Kemeny, $\mathcal{O}(n^3)$ for Hamming (n =number of labels)

Theoretical and computational advantages

For Kemeny and Hamming embedding:

- ▶ **consistency holds:**

$$\mathcal{R}(d \circ \hat{g}) - \mathcal{R}(s^*) \leq c_\phi \sqrt{\mathcal{L}(\hat{g}) - \mathcal{L}(g^*)}$$

with $c_{\phi_\tau} = \sqrt{\frac{n(n-1)}{2}}$ and $c_{\phi_H} = \sqrt{n}$ (constants with n)

- ▶ but the **pre-image step is hard**: NP-hard for Kemeny, $\mathcal{O}(n^3)$ for Hamming (n =number of labels)

In contrast, for the Lehmer embedding:

- ▶ we **lose consistency**:

$$\mathcal{R}(d \circ \hat{g}) - \mathcal{R}(s^*) \leq \sqrt{\frac{n(n-1)}{2}} \sqrt{\mathcal{L}(\hat{g}) - \mathcal{L}(g^*)} + \mathcal{R}(d \circ g^*) - \mathcal{R}(s^*)$$

- ▶ but the **pre-image step is simple**: $\mathcal{O}(n)$

Structured prediction - Numerical results

Table: Mean Kendall's τ coefficient on benchmark datasets

	authorship	glass	iris	vehicle	vowel	wine
kNN Hamming	0.01±0.02	0.08±0.04	-0.15±0.13	-0.21±0.04	0.24±0.04	-0.36±0.04
kNN Kemeny	0.94 ±0.02	0.85±0.06	0.95±0.05	0.85±0.03	0.85±0.02	0.94±0.05
kNN Lehmer	0.93±0.02	0.85±0.05	0.95±0.04	0.84±0.03	0.78±0.03	0.94±0.06
ridge Hamming	-0.00±0.02	0.08±0.05	-0.10±0.13	-0.21±0.03	0.26±0.04	-0.36±0.03
ridge Lehmer	0.92±0.02	0.83±0.05	0.97 ±0.03	0.85±0.02	0.86±0.01	0.84±0.08
ridge Kemeny	0.94 ±0.02	0.86±0.06	0.97 ±0.05	0.89 ±0.03	0.92 ±0.01	0.94±0.05
Cheng PL	0.94 ±0.02	0.84±0.07	0.96±0.04	0.86±0.03	0.85±0.02	0.95 ±0.05
Cheng LWD	0.93±0.02	0.84±0.08	0.96±0.04	0.85±0.03	0.88±0.02	0.94±0.05
Zhou RF	0.91	0.89	0.97	0.86	0.87	0.95

Cheng PL [Cheng et al., 2010], Cheng LWD [Cheng and Hüllermeier, 2013], Zhou RF [Zhou and Qiu, 2016]

Kendall's τ coefficient corresponds to a rescaling of Kendall's tau distance d_τ between [-1,1] (so the closer from 1 is the better)

Outline

Introduction to Ranking Data

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

Ranking Aggregation - A Statistical Framework

$$\min_{\sigma \in \mathfrak{S}_n} L(\sigma), \text{ with } L(\sigma) = \mathbb{E}[d_\tau(\sigma, \Sigma)]$$

Label Ranking - Piecewise Constant Rules

$$\min_{s : \mathcal{X} \rightarrow \mathfrak{S}_n} \mathcal{R}(s), \text{ with } \mathcal{R}(s) = \mathbb{E}[d_\tau(s(X), \Sigma)]$$

Label Ranking - A Structured Prediction Approach

$$\min_{s : \mathcal{X} \rightarrow \mathfrak{S}_n} \mathcal{R}(s), \text{ with } \mathcal{R}(s) = \mathbb{E}[\Delta(s(X), \Sigma)]$$

Conclusion

Other contributions

A general method to bound the distance of a ranking to a Kemeny consensus [Jiao et al., 2016]

Given:

- ▶ a dataset $D_N = (\sigma_1, \dots, \sigma_N)$
 - ▶ a ranking σ (typically the output of a tractable voting rule)
- $$\Rightarrow d(\sigma, \sigma_{D_N}^*) \leq k \text{ where } k = k(\sigma, D_N) \in \{0, \dots, \binom{n}{2}\}$$

Dimensionality reduction for ranking data [Achab et al., 2018]

How to approximate P on \mathfrak{S}_n by a \tilde{P} parameterized by much less parameters (i.e. $\ll n!$)?

$$\tilde{P}_{\mathcal{C}}^* \in \min_{P' \in \mathbf{P}_{\mathcal{C}}} W_{d_{\tau}, 1}(P, P').$$

where \mathcal{C} is a bucket order/partial ranking and $\mathbf{P}_{\mathcal{C}}$ is the set of distributions such that:

$$P' \in \mathbf{P}_{\mathcal{C}} \Rightarrow P'(\sigma) = 0 \text{ iff } \sigma \text{ disagree with } \mathcal{C}$$

Conclusion

Ranking data presents great and interesting challenges:

- ▶ Most of the maths from euclidean spaces cannot be applied
- ▶ But our intuitions still hold
- ▶ Our contributions: theoretical results for ranking aggregation, label ranking, and new algorithms

Openings/Perspectives:

- ▶ How to extend to incomplete rankings, with ties (as input or output)?
- ▶ How to weight specific items/ranks?

-  Achab, M., Korba, A., and Cléménçon, S. (2018).
Dimensionality reduction and (bucket) ranking: a mass transportation approach.
arXiv preprint arXiv:1810.06291.
-  Alvo, M. and Yu, P. L. H. (2014).
Statistical Methods for Ranking Data.
Springer.
-  Audibert, J. Y. and Tsybakov, A. (2007).
Fast learning rates for plug-in classifiers.
Annals of statistics, 35(2):608–633.
-  Borda, J. C. (1781).
Mémoire sur les élections au scrutin.
-  Brouard, C., Szafranski, M., and d'Alché Buc, F. (2016).
Input output kernel regression: supervised and
semi-supervised structured output prediction with
operator-valued kernels.
Journal of Machine Learning Research, 17(176):1–48.

-  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.
-  Cheng, W. and Hüllermeier, E. (2013).
A nearest neighbor approach to label ranking based on
generalized labelwise loss minimization.
-  Cheng, W., Hüllermeier, E., and Dembczynski, K. J. (2010).
Label ranking methods based on the plackett-luce model.
In *Proceedings of the 27th International Conference on Machine
Learning (ICML)*, pages 215–222.
-  Ciliberto, C., Rosasco, L., and Rudi, A. (2016).
A consistent regularization approach for structured prediction.
In *Advances in Neural Information Processing Systems (NIPS)*,
pages 4412–4420.
-  Clémenton, S., Gaudel, R., and Jakubowicz, J. (2011).

Clustering rankings in the fourier domain.
In *Machine Learning and Knowledge Discovery in Databases*,
pages 343–358. Springer.

-  Clémençon, S., Korba, A., and Sibony, E. (2018).
Ranking median regression: Learning to order through local
consensus.
International Conference on Algorithmic Learning Theory (ALT).

-  Copeland, A. H. (1951).
A reasonable social welfare function.
In *Seminar on applications of mathematics to social sciences*,
University of Michigan.

-  Davidson, D. and Marschak, J. (1959).
Experimental tests of a stochastic decision theory.
Measurement: Definitions and theories, 17:274.

-  Dwork, C., Kumar, R., Naor, M., and Sivakumar, D. (2001).
Rank aggregation methods for the web.

In *Proceedings of the 10th International Conference on World Wide Web*, pages 613–622. ACM.

-  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 *Proceedings of the 33rd International Conference on Machine Learning (ICML)*.
-  Jiao, Y. and Vert, J. (2015).
The kendall and mallows kernels for permutations.
In Blei, D. and Bach, F., editors, *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 1935–1944.
-  Jiao, Y. and Vert, J.-P. (2018).
The weighted Kendall and high-order kernels for permutations.

In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 2319–2327, Stockholmsmässan, Stockholm Sweden. PMLR.

-  Kemeny, J. G. (1959).
Mathematics without numbers.
Daedalus, 88:571–591.
-  Kondor, R. and Barbosa, M. S. (2010).
Ranking with kernels in Fourier space.
In *The 23rd Conference on Learning Theory (COLT)*, pages 451–463.
-  Korba, A., Clémenton, S., and Sibony, E. (2017).
A learning theory of ranking aggregation.
In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*.
-  Korba, A., Garcia, A., and Buc d'Alché, F. (2018).
A structured prediction approach for label ranking.

-  Li, P., Mazumdar, A., and Milenkovic, O. (2017).
Efficient rank aggregation via lehmer codes.
Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS).
-  Lu, T. and Boutilier, C. (2014).
Effective sampling and learning for mallows models with pairwise-preference data.
volume 15, pages 3963–4009.
-  Luce, R. D. (1959).
Individual Choice Behavior.
Wiley.
-  Mallows, C. L. (1957).
Non-null ranking models.
Biometrika, 44(1-2):114–130.
-  Plis, S., McCracken, S., Lane, T., and Calhoun, V. (2011).

Directional statistics on permutations.

In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 600–608.



Rajkumar, A. and Agarwal, S. (2014).

A statistical convergence perspective of algorithms for rank aggregation from pairwise data.

In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 118–126.



Shah, N. B. and Wainwright, M. J. (2017).

Simple, robust and optimal ranking from pairwise comparisons.

Journal of Machine Learning Research.



Sibony, E., Clemençon, S., and Jakubowicz, J. (2015).

Mra-based statistical learning from incomplete rankings.

In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 1432–1441.



Sun, M., Lebanon, G., and Kidwell, P. (2012).

Estimating probabilities in recommendation systems.

Journal of the Royal Statistical Society: Series C (Applied Statistics), 61(3):471–492.



Szörényi, B., Busa-Fekete, R., Paul, A., and Hüllermeier, E. (2015).

Online rank elicitation for plackett-luce: A dueling bandits approach.

In *Advances in Neural Information Processing Systems (NIPS)*, pages 604–612.



Tversky, A. (1972).

Elimination by aspects: A theory of choice.

Psychological review, 79(4):281.



Zhao, Z., Piech, P., and Xia, L. (2016).

Learning mixtures of plackett-luce models.

In *Proceedings of the 33nd International Conference on Machine Learning (ICML)*, pages 2906–2914.



Zhou, Y. and Qiu, G. (2016).

Random forest for label ranking.

arXiv preprint arXiv:1608.07710.

A general method to bound the distance to Kemeny consensus

For all $\sigma, \sigma' \in \mathfrak{S}_n$, $\|\phi(\sigma) - \phi(\sigma')\|^2 = 4d_\tau(\sigma, \sigma')$ where $\phi = \phi_\tau$.

Thus solving Kemeny's rule on $\mathcal{D}_N = (\sigma_1, \dots, \sigma_N) \Leftrightarrow$ solving:

$$\min_{\sigma \in \mathfrak{S}_n} \|\phi(\sigma) - \phi(\mathcal{D}_N)\|^2$$

where $\phi(\mathcal{D}_N) := \frac{1}{N} \sum_{t=1}^N \phi(\sigma_t)$.

For any $\sigma \in \mathfrak{S}_n$:

- ▶ $\cos(\theta_N(\sigma)) = \frac{\langle \phi(\sigma), \phi(\mathcal{D}_N) \rangle}{\|\phi(\sigma)\| \|\phi(\mathcal{D}_N)\|}$
- ▶ $k_{min}(\sigma; \mathcal{D}_N) = \begin{cases} \left\lfloor \binom{n}{2} \sin^2(\theta_N(\sigma)) \right\rfloor & \text{if } 0 \leq \theta_N(\sigma) \leq \frac{\pi}{2} \\ \binom{n}{2} & \text{if } \frac{\pi}{2} \leq \theta_N(\sigma) \leq \pi. \end{cases}$

Dimensionality reduction for ranking data: a mass transportation approach

A bucket order $\mathcal{C} = (\mathcal{C}_1, \dots, \mathcal{C}_K)$ is a partial ranking, i.e.:

- ▶ $\{\mathcal{C}_k\}_{1 \leq k \leq K}$ is a partition of $\llbracket n \rrbracket$
- ▶ $\mathcal{C}_1 \prec \dots \prec \mathcal{C}_K$

K : size of the bucket, $\lambda = (\#\mathcal{C}_1, \dots, \#\mathcal{C}_K)$: shape of the bucket (verifying $\sum_{k=1}^K \#\mathcal{C}_k = n$).

How to approximate P on \mathfrak{S}_n by a \tilde{P} parameterized by much less parameters (i.e. $\ll n!$)?

We propose to represent P by a bucket order \mathcal{C} and an approximate distribution $\tilde{P}_{\mathcal{C}}^*$ relative to \mathcal{C} :

$$\tilde{P}_{\mathcal{C}}^* \in \min_{P' \in \mathbf{P}_{\mathcal{C}}} W_{d_{\tau}, 1}(P, P').$$

where $\mathbf{P}_{\mathcal{C}}$ is the set of bucket distributions:

$P' \in \mathbf{P}_{\mathcal{C}} \Rightarrow$ for all $1 \leq k < l \leq K$ and all $(i, j) \in \llbracket K \rrbracket^2$,

$$(i, j) \in \mathcal{C}_k \times \mathcal{C}_l \Rightarrow p'_{i,j} = \mathbb{P}_{\Sigma' \sim P'}[\Sigma'(i) < \Sigma'(j)] = 0 \quad 46/47$$

Dimensionality reduction for ranking data: a mass transportation approach

$$\Lambda_P(\mathcal{C}) = \min_{P' \in \mathbf{P}_{\mathcal{C}}} W_{d_{\tau},1}(P, P') = \sum_{i \prec_{\mathcal{C}} j} p_{j,i}.$$

We denote by \mathbf{C}_K the set of all bucket orders \mathcal{C} of $[\![n]\!]$ with K buckets.

Our proposal: find a solution $\mathcal{C}^{*(K)}$ of

$$\min_{\mathcal{C} \in \mathbf{C}_K} \Lambda_P(\mathcal{C}), \tag{3}$$