Controlling the distance to the Kemeny consensus without computing it

Eric Sibony Yunlong Jiao Anna Korba

LTCI UMR 5141, Telecom ParisTech/CNRS, Mines ParisTech

ICML 2016

Outline

The ranking aggregation problem can be encoutered in many fields of the scientific literature

- Elections in Social choice theory
- Meta search engines
- Competitions rankings
- Analysis of biological data
- Natural Language Processing

Ranking aggregation

Problem:

How to summarize a collection of rankings into one ranking?

Input

- Set of items: $[\![n]\!] := \{1, \ldots, n\}$
- N Rankings of the form : $i_1 \succ \cdots \succ i_n$

Output

A global order ("consensus") σ^* on the n objects.

Ranking aggregation

Ranking $i_1 \succ \cdots \succ i_n$ on $[\![n]\!]$ $\iff$ permutation σ on $[\![n]\!]$ s.t. $\sigma(i_j) = j$.

Ranking aggregation

Ranking $i_1 \succ \cdots \succ i_n$ on $[\![n]\!]$ $\iff$ permutation σ on $[\![n]\!]$ s.t. $\sigma(i_j) = j$.

What permutation $\sigma^* \in \mathfrak{S}_n$ best represents a given a collection of permutations $(\sigma_1, \ldots, \sigma_N) \in \mathfrak{S}_n^N$?

Ranking aggregation

Ranking $i_1 \succ \cdots \succ i_n$ on $[\![n]\!]$ $\iff$ permutation σ on $[\![n]\!]$ s.t. $\sigma(i_j) = j$.

What permutation $\sigma^* \in \mathfrak{S}_n$ best represents a given a collection of permutations $(\sigma_1, \ldots, \sigma_N) \in \mathfrak{S}_n^N$?

Definition (*Consensus ranking (Kemeny, 1959)*)

A permutation $\sigma^* \in \mathfrak{S}_n$ is a best representative of the collection $(\sigma_1, \ldots, \sigma_N) \in \mathfrak{S}_n^N$ with respect to a metric d on $\mathfrak{S}_n$ if it is a solution of :

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

Kemeny's rule

Definition (*Kendall's distance*)

The Kendalls tau distance between two permutations is equal to the number of their pairwise disagreements:

$$d_{KT}(\sigma, \pi) = \sum_{\{i,j\} \subset [\![n]\!]} \mathbb{I}\{\sigma \textit{ and } \pi \textit{ disagree on } \{i,j\}\}$$

Example

σ= 123 ($1 \succ 2 \succ 3$)
π= 231 ($2 \succ 3 \succ 1$)
$\rightarrow$ number of desagreements = on 2 pairs (12,13).

Kemeny's rule

Definition (*Kendall's distance*)

The Kendalls tau distance between two permutations is equal to the number of their pairwise disagreements:

$$d_{KT}(\sigma, \pi) = \sum_{\{i,j\} \subset [\![n]\!]} \mathbb{I}\{\sigma \text{ and } \pi \text{ disagree on } \{i,j\}\}$$

Example

$\sigma = 123 \ (1 \succ 2 \succ 3)$
$\pi = 231 \ (2 \succ 3 \succ 1)$
$\rightarrow$ number of desagreements = on 2 pairs (12,13).

Definition (*Kemeny's rule*)

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

Kemeny's rule

- Social choice justification: Satisfies many voting properties, such as the Condorcet criterion: if a candidate 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: It is NP-hard in the number of votes N [Bartholdi et al., 1989] even for $n = 4$ candidates [Dwork et al., 2001].

Outline

Contribution

Previous contributions

- General guarantees for approximation procedures
- Bounds on the approximation cost of procedures
- Conditions for the exact Kemeny aggregation to become tractable

Contribution

Previous contributions

- General guarantees for approximation procedures
- Bounds on the approximation cost of procedures
- Conditions for the exact Kemeny aggregation to become tractable

Our approach

- Set of items $[\![n]\!] := \{1, \ldots, n\}$
- A rankings dataset $\mathcal{D}_N = (\sigma_1, \ldots, \sigma_N) \in \mathfrak{S}_n^N$
- Let $\sigma \in \mathfrak{S}_n$ a permutation, typically out put by a computationally efficient aggregation procedure on $\mathcal{D}_N$.

Can we give an upper bound $d(\sigma, \sigma^*)$ between σ and a Kemeny consensus, by using only tractable quantities?

Outline

Kemeny embedding

The Kemeny embedding is the mapping $\phi : \mathfrak{S}_n \to \mathbb{R}^{\binom{n}{2}}$ defined by:

$$\phi : \sigma \mapsto \begin{pmatrix} \vdots \\ sign(\sigma(i) - \sigma(j)) \\ \vdots \end{pmatrix}_{1 \leq i < j \leq n}$$

where $sign(x) = 1$ if $x \geq 0$ and 1 otherwise.

Example

$$123 \mapsto \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \begin{matrix} \to \text{pair } 12 \\ \to \text{pair } 13 \\ \to \text{pair } 23 \end{matrix} \quad , 132 \mapsto \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} \begin{matrix} \to \text{pair } 12 \\ \to \text{pair } 13 \\ \to \text{pair } 23 \end{matrix}$$

Kemeny aggregation in $\mathbb{R}^{\binom{n}{2}}$

Definition (*Mean embedding*)

For $D_N = (\sigma_1, \ldots, \sigma_N) \in \mathfrak{S}_n^N$, *we define the* **barycenter***:*

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

Kemeny aggregation in $\mathbb{R}^{\binom{n}{2}}$

Definition (*Mean embedding*)

For $D_N = (\sigma_1, \ldots, \sigma_N) \in \mathfrak{S}_n^N$, we define the **barycenter***:*

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

Proposition (*Barthelemy & Monjardet (1981)*)

For all $\sigma, \sigma' \in \mathfrak{S}_n$,

$$\|\phi(\sigma)\| = \sqrt{\frac{n(n-1)}{2}} \quad \text{and} \quad \|\phi(\sigma) - \phi(\sigma')\|^2 = 4d(\sigma, \sigma'),$$

and for any dataset $\mathcal{D}_N = (\sigma_1, \ldots \sigma_N) \in \mathfrak{S}_n^N$, Kemeny aggregation (1) is equivalent to the minimization problem

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

Illustration

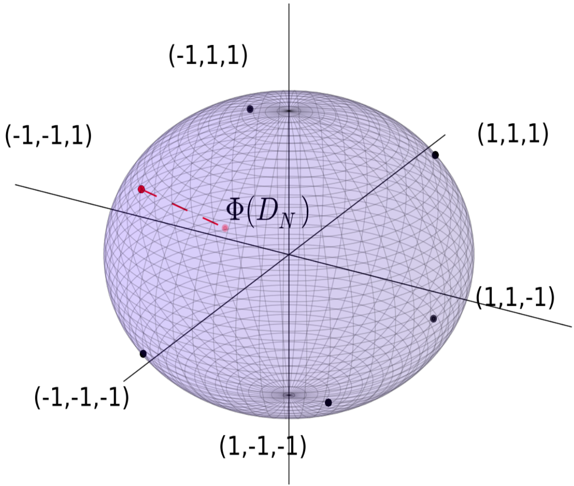

Figure: Kemeny aggregation for $n = 3$.

Kemeny aggregation in $\mathbb{R}^{\binom{n}{2}}$

Kemeny aggregation naturally decomposes in two steps:

1. Compute the barycenter $\phi(\mathcal{D}_N) \in \mathbb{R}^{\binom{n}{2}}$ (complexity $O(Nn^2)$)
2. Find the consensus σ^* solution of problem (2)

Main result

For $\sigma \in \mathfrak{S}_n$, we define the angle $\theta_N(\sigma)$ between $\phi(\sigma)$ and $\phi(\mathcal{D}_N)$ by:

$$\cos(\theta_N(\sigma)) = \frac{\langle \phi(\sigma), \phi(\mathcal{D}_N) \rangle}{\|\phi(\sigma)\| \|\phi(\mathcal{D}_N)\|},$$

with $0 \leq \theta_N(\sigma) \leq \pi$.

Main result

For $\sigma \in \mathfrak{S}_n$, we define the angle $\theta_N(\sigma)$ between $\phi(\sigma)$ and $\phi(\mathcal{D}_N)$ by:

$$\cos(\theta_N(\sigma)) = \frac{\langle \phi(\sigma), \phi(\mathcal{D}_N) \rangle}{\|\phi(\sigma)\| \|\phi(\mathcal{D}_N)\|},$$

with $0 \leq \theta_N(\sigma) \leq \pi$.

Theorem

Let $\mathcal{D}_N \in \mathfrak{S}_n^N$ be a dataset, $\mathcal{K}_N$ the set of Kemeny consensuses and $\sigma \in \mathfrak{S}_n$ a permutation. For any $k \in \{0, \dots, \binom{n}{2} - 1\}$, one has the following implication:

$$\cos(\theta_N(\sigma)) > \sqrt{1 - \frac{k+1}{\binom{n}{2}}} \quad \Rightarrow \quad \max_{\sigma^* \in \mathcal{K}_N} d(\sigma, \sigma^*) \leq k.$$

Method

We define:

$$k_{min}(\sigma; \mathcal{D}_N) = \left\lfloor \binom{n}{2} \sin^2(\theta_N(\sigma)) \right\rfloor . \tag{3}$$

the minimal $k \in \{0, \ldots, \binom{n}{2} - 1\}$ verifying the theorem condition.

Two steps:

- Compute $k_{min}(\sigma; \mathcal{D}_N)$ with Formula (3).
- Then by Theorem 15, $d(\sigma, \sigma^*) \leq k_{min}(\sigma; \mathcal{D}_N)$ for all Kemeny consenus $\sigma^* \in \mathcal{K}_N$.

Application on the sushi dataset

Table: Summary of a case-study on the validity of the method with the sushi dataset ($N = 5000, n = 10$). Rows are ordered by increasing k_{min} (or decreasing cosine) value.

Voting rule	$\cos(\theta_N(\sigma))$	$k_{min}(\sigma)$
Borda	0.82	14
Copeland	0.82	14
QuickSort	0.82	14
Plackett-Luce	0.80	15
2-approval	0.74	20
1-approval	0.71	22
Pick-a-Perm	0.40	37
Pick-a-Random	0.28	41

Outline

Extended cost function

Kemeny aggregation:

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

Relaxed problem:

$$\min_{x \in \mathbb{S}} \mathcal{C}_N(x) := \|x - \phi(\mathcal{D}_N)\|^2. \tag{4}$$

Extended cost function

Kemeny aggregation:

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

Relaxed problem:

$$\min_{x \in \mathbb{S}} \mathcal{C}_N(x) := \|x - \phi(\mathcal{D}_N)\|^2. \tag{4}$$

For any $x \in \mathbb{S}$, by denoting R the radius of $\mathbb{S}$, one has:

$$\mathcal{C}_N(x) = R^2 + \|\phi(\mathcal{D}_N)\|^2 - 2R\|\phi(\mathcal{D}_N)\| \cos(\theta_N(x)).$$

The level sets of $\mathcal{C}_N$ are thus of the form $\{x \in \mathbb{S} \mid \theta_N(x) = \alpha\}$, for $0 \leq \alpha \leq \pi$

Illustration

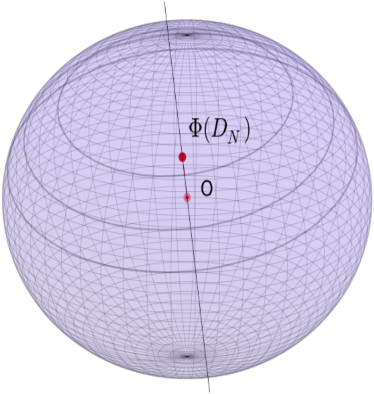

Figure: Level sets of $\mathcal{C}_N$

Lemmas

Lemma (1)

A Kemeny consensus of a dataset $\mathcal{D}_N$ is a permutation σ^* s.t:

$$\theta_N(\sigma^*) \leq \theta_N(\sigma) \qquad \textit{for all } \sigma \in \mathfrak{S}_n.$$

We denote by $\mathcal{B}(x, r) = \{x' \in \mathbb{R}^{\binom{n}{2}} \mid \|x' - x\| < r\}$ the (open) ball of center x and radius r.

Lemma (2)

For $x \in \mathbb{S}$ and $r \geq 0$, one has:

$$\cos(\theta_N(x)) > \sqrt{1 - \frac{r^2}{4R^2}} \Rightarrow \min_{x' \in \mathbb{S} \setminus \mathcal{B}(x,r)} \theta_N(x') > \theta_N(x).$$

Illustration

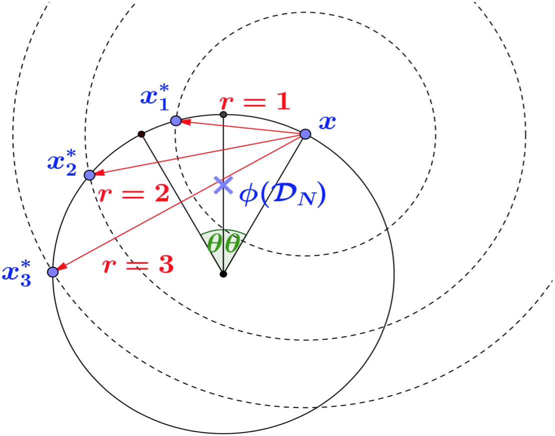

Figure: Illustration of Lemma 2 with r taking integer values (representing possible Kendall's tau distance).

Embedding of a ball

Lemma (3)

For $\sigma \in \mathfrak{S}_n$ *and* $k \in \{0, \ldots, \binom{n}{2}\}$,

$$\phi\left(\mathfrak{S}_n \setminus B(\sigma, k)\right) \quad \subset \quad \mathbb{S} \setminus \mathcal{B}(\phi(\sigma), 2\sqrt{k+1})$$

Outline

Applicability of the method

We denote by:

- n the number of alternatives
- $\mathcal{D}_N \in \mathfrak{S}_n^N$ any dataset
- r any voting rule, and by $r(\mathcal{D}_N)$ the consensuses of $\mathcal{D}_N$ given by r

We know that:

$$d(r(\mathcal{D}_N), \mathcal{K}_N) \leq k_{min}\,.$$

We study the tightness of the bound:

$$s\left(r, \mathcal{D}_N, n\right) := k_{min} - d(r(\mathcal{D}_N), \mathcal{K}_N)\,.$$

Results

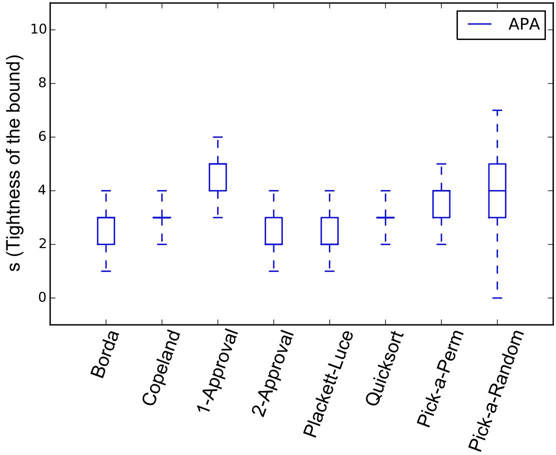

Figure: Boxplot of $s\left(r, \mathcal{D}_N, n\right)$ over sampling collections of datasets shows the effect from different voting rules r with 500 bootstrapped pseudo-samples of the APA dataset ($n = 5$, $N = 5738$).

Predictability of the method

- When n grows, the exact Kemeny consensus $\mathcal{K}_N$, hence $s\left(r, \mathcal{D}_N, n\right)$ quickly becomes computationally impermissible.

- Once we have an approximate ranking $r(\mathcal{D}_N)$ and k_{min} is identified via our method, the search scope for the exact Kemeny consensuses can be narrowed down to those permutations within a distance of k_{min} to $r(\mathcal{D}_N)$.

- Notably the total number of such permutations in $\mathfrak{S}_n$ is upper bounded by $\binom{n+k_{min}-1}{k_{min}} << |\mathfrak{S}_n| = n!$ [Wang 2013].

Results

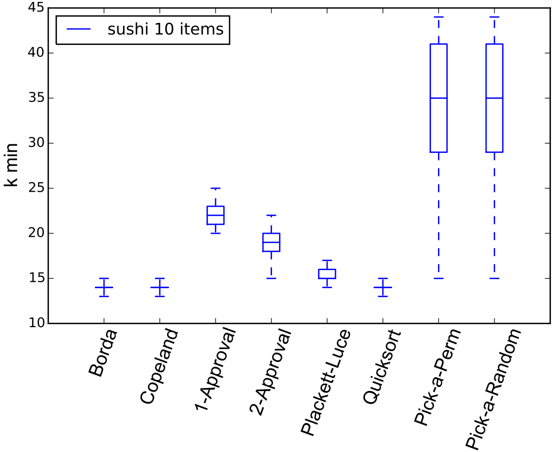

Figure: Boxplot of k_{min} over 500 bootstrapped pseudo-samples of the sushi dataset ($n = 10, N = 5000$).

Outline

Conclusion

- We have established a theoretical result that allows to control the Kendall's tau distance between a permutation and the Kemeny consensuses of any dataset.

Conclusion

- We have established a theoretical result that allows to control the Kendall's tau distance between a permutation and the Kemeny consensuses of any dataset.
- This provides a simple and general method to predict, for any ranking aggregation procedure, how close the outcome on a dataset is from the Kemeny consensuses.

Future directions

- The geometric properties of the Kemeny embedding are rich and could lead to many more results.

Future directions

- The geometric properties of the Kemeny embedding are rich and could lead to many more results.
- We can imagine ranking aggregation procedures using a smaller scope for Kemeny consensuses.

Future directions

- The geometric properties of the Kemeny embedding are rich and could lead to many more results.
- We can imagine ranking aggregation procedures using a smaller scope for Kemeny consensuses.
- Possible extensions to incomplete rankings.

Thank you