Consensus Ranking Distribution as Ranking Quantization: A k-medians Algorithm

Gabriel Benedetti   Stéphan Clémençon   Ekhine Irurozki

LTCI, Télécom Paris, Institut Polytechnique de Paris

DA2PL

Outline

1. Motivation
2. Background
3. Consensus Ranking Distributions
4. The k-medians algorithms
5. Experiments

Outline

1. Motivation
2. Background
3. Consensus Ranking Distributions
4. The k-medians algorithms
5. Experiments

Rankings - mixtures of well concentrated components

Ranking data is ubiquitous:

- **Recommendation systems**: users rank products; preferences cluster into groups.

Rankings - mixtures of well concentrated components

Ranking data is ubiquitous:

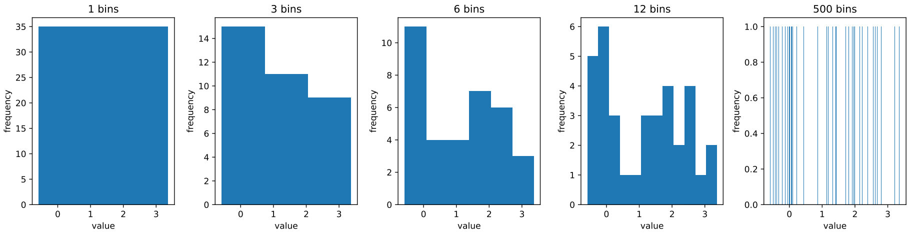
- **Recommendation systems**: users rank products; preferences cluster into groups.
- **Elections & voting**: electorates split into factions with distinct preference orders.
- **Benchmarking**: tasks/instances cluster the algorithms according to their performance
- **Information retrieval**: search engine results ranked by multiple criteria.

A single Kemeny median fails when P is:

- **Heterogeneous**: different sub-populations with distinct ranking patterns.
- **High-variance**: large spread around any single central ranking.

We need a *richer* summary — one that adapts to the structure of P.

Motivation: quantizing a distribution



Histograms approximate a distribution with increasing resolution as the number of bins K grows.

- $K = 1$: one bin = one mean/median.
  Maximum compression, maximum distortion.
- K **larger**: finer approximation.
  Reveals modes, tails, heterogeneity.
- K **too large**: overfitting to the sample.

Our question

Can we do the same for distributions over **rankings**?

$\Rightarrow$ **Consensus Ranking Distribution (CRD)**

From histograms to ranking distributions

Histogram of P on $\mathbb{R}$:

- Partition $\mathbb{R}$ into K bins $\mathcal{C}_1, \ldots, \mathcal{C}_K$.
- Represent each bin by its *local mean* μ_k.
- Weight $= P(\mathcal{C}_k)$.
- Approximation: $\widehat{P} = \sum_k P(\mathcal{C}_k)\, \delta_{\mu_k}$.

CRD of P on $\mathfrak{S}_n$:

- Partition $\mathfrak{S}_n$ into K cells $\mathcal{C}_1, \ldots, \mathcal{C}_K$.
- Represent each cell by its *local Kemeny median* σ_k^*.
- Weight $= P(\mathcal{C}_k)$.
- $P_{\mathcal{P}} = \sum_k P(\mathcal{C}_k)\, \delta_{\sigma_k^*}$.

Key challenges specific to $\mathfrak{S}_n$

- No vector space structure $\Rightarrow$ no mean; use *Kendall median* instead.
- $|\mathfrak{S}_n| = n!$ is astronomically large; $N \ll n!$ in practice.
- The right notion of "bin" on rankings must be designed carefully.

Outline

1. Motivation
2. Background
3. Consensus Ranking Distributions
4. The k-medians algorithms
5. Experiments

Setup: ranking distributions on $\mathfrak{S}_n$

- n items indexed by $\{1, \ldots, n\}$.
- A **ranking**: ranking $\sigma \in \mathfrak{S}_n$, $\sigma(i) =$ rank of item i.
- **Kendall τ distance** — discordant pairs:

$$d_\tau(\sigma, \sigma') = \sum_{i<j} 1\left[(\sigma(i) - \sigma(j))(\sigma'(i) - \sigma'(j)) < 0\right].$$

- **Pairwise marginals**: $p_{ij} = \mathbb{P}\{\Sigma(i) < \Sigma(j)\}$, $p_{ij} + p_{ji} = 1$.
  Do not fully characterise P in general

Setup: ranking distributions on $\mathfrak{S}_n$

- n items indexed by $\{1, \dots, n\}$.
- A **ranking**: ranking $\sigma \in \mathfrak{S}_n$, $\sigma(i)$ = rank of item i.
- **Kendall τ distance** — discordant pairs:

$$d_\tau(\sigma, \sigma') = \sum_{i<j} \mathbb{1}\big[(\sigma(i)-\sigma(j))(\sigma'(i)-\sigma'(j)) < 0\big].$$

- **Pairwise marginals**: $p_{ij} = \mathbb{P}\{\Sigma(i) < \Sigma(j)\}$, $p_{ij} + p_{ji} = 1$.
  Do not fully characterise P in general

Ranking risk:

$$L_P(\sigma) = \mathbb{E}_{\Sigma \sim P}[d_\tau(\Sigma, \sigma)]$$

Variability measures:

$$V_P = \min_\sigma L_P(\sigma) = \min_\sigma \mathbb{E}_{\Sigma \sim P}[d_\tau(\Sigma, \sigma)],$$

$$V'_P = \tfrac{1}{2}\mathbb{E}[d_\tau(\Sigma, \Sigma')]$$

Setup: ranking distributions on $\mathfrak{S}_n$

- n items indexed by $\{1, \ldots, n\}$.
- A **ranking**: ranking $\sigma \in \mathfrak{S}_n$, $\sigma(i) =$ rank of item i.
- **Kendall τ distance** — discordant pairs:

$$d_\tau(\sigma, \sigma') = \sum_{i<j} 1\left[(\sigma(i)-\sigma(j))(\sigma'(i)-\sigma'(j)) < 0\right].$$

- **Pairwise marginals**: $p_{ij} = \mathbb{P}\{\Sigma(i) < \Sigma(j)\}$, $p_{ij} + p_{ji} = 1$.
  Do not fully characterise P in general

Ranking risk:

$$L_P(\sigma) = \mathbb{E}_{\Sigma \sim P}[d_\tau(\Sigma, \sigma)]$$

Variability measures:

$$V_P = \min_\sigma L_P(\sigma) = \min_\sigma \mathbb{E}_{\Sigma \sim P}[d_\tau(\Sigma, \sigma)],$$

$$V'_P = \tfrac{1}{2}\mathbb{E}[d_\tau(\Sigma, \Sigma')]$$

The median minimizes V_P.

$$\boxed{V'_P \leq V_P \leq 2V'_P}$$

Consensus ranking and Kemeny medians

The **Kemeny median in CO** is the ranking

$$\sigma^* \in \arg\min_{\sigma \in \mathfrak{S}_n} \sum_{\sigma' \in S} d_\tau(\sigma, \sigma'). \qquad \text{(NP-hard)}$$

The **Kemeny median** is the ranking that minimises risk:

$$\sigma^* \in \arg\min_{\sigma \in \mathfrak{S}_n} L_P(\sigma) = \arg\min_{\sigma \in \mathfrak{S}_n} \mathbb{E}[d_\tau(\Sigma, \sigma)]. \qquad \text{(NP-hard)}$$

Optimal transport view. Define the Wasserstein metric:

$$W_d(P, P') = \inf_{(\Sigma, \Sigma') \sim (P, P')} \mathbb{E}[d(\Sigma, \Sigma')].$$

Then: $\sigma^* = \arg\min_\sigma W_{d_\tau}(P, \delta_\sigma)$.

Kemeny $\equiv$ is the Dirac mass that is closest to P in Wasserstein distance.

Consensus ranking and Kemeny medians

The **Kemeny median in CO** is the ranking

$$\sigma^* \in \arg\min_{\sigma \in \mathfrak{S}_n} \sum_{\sigma' \in S} d_\tau(\sigma, \sigma'). \qquad \text{(NP-hard)}$$

The **Kemeny median** is the ranking that minimises risk:

$$\sigma^* \in \arg\min_{\sigma \in \mathfrak{S}_n} L_P(\sigma) = \arg\min_{\sigma \in \mathfrak{S}_n} \mathbb{E}[d_\tau(\Sigma, \sigma)]. \qquad \text{(NP-hard)}$$

Optimal transport view. Define the Wasserstein metric:

$$W_d(P, P') = \inf_{(\Sigma, \Sigma') \sim (P, P')} \mathbb{E}[d(\Sigma, \Sigma')].$$

Then: $\sigma^* = \arg\min_\sigma W_{d_\tau}(P, \delta_\sigma)$.

Kemeny $\equiv$ is the Dirac mass that is closest to P in Wasserstein distance.

Our proposal: instead of one Dirac mass, use K of them. That is the CRD.

Outline

1. Motivation
2. Background
3. Consensus Ranking Distributions
4. The k-medians algorithms
5. Experiments

Local ranking medians

Definition (Local consensus)

Let $\mathcal{C} \subset \mathfrak{S}_n$ with $P(\mathcal{C}) > 0$. Denote the conditional distribution of Σ given $\Sigma \in \mathcal{C}$ by $P_{\mathcal{C}}$. A **local median** $\sigma^*_{\mathcal{C}}$ satisfies:

$$\mathbb{E}[d_\tau(\Sigma, \sigma^*_{\mathcal{C}}) \mid \Sigma \in \mathcal{C}] = V(\mathcal{C}) := \min_{\sigma \in \mathfrak{S}_n} L_{P_{\mathcal{C}}}(\sigma).$$

The local variability measures for a cell $\mathcal{C}$:

$$V(\mathcal{C}) = \min_{\sigma} \mathbb{E}[d_\tau(\Sigma, \sigma) \mid \Sigma \in \mathcal{C}] \qquad \text{(Kendall distortion, NP-hard)}$$

$$V'(\mathcal{C}) = \mathbb{E}[d_\tau(\Sigma, \Sigma') \mid \Sigma \in \mathcal{C}] \qquad \text{(always computable, lower bound)}$$

where $V(\mathcal{C}) \leq V'(\mathcal{C})$.

Key subtlety

$\sigma^*_{\mathcal{C}}$ need **not** belong to $\mathcal{C}$ — it is the median of the conditional distribution.

Definition: Consensus Ranking Distribution (CRD)

Definition (CRD)

Let $\mathcal{P} = \{\mathcal{C}_1, \dots, \mathcal{C}_K\}$ be a partition of $\mathfrak{S}_n$. The **consensus ranking distribution** of P w.r.t. $\mathcal{P}$ is:

$$\boxed{P_{\mathcal{P}} = \sum_{\mathcal{C} \in \mathcal{P}} P(\mathcal{C})\, \delta_{\sigma^*_{\mathcal{C}}}.}$$

The spectrum of approximations:

$K = 1$	K cells	$K = n!$	
δ_{σ^*}	$P_{\mathcal{P}}$	$\widehat{P}_N$	
Kemeny	CRD	empirical	$\longrightarrow$ complexity

distortion $\downarrow$ complexity $\uparrow$

Distortion bound (Proposition 1)

Proposition 1

For any partition $\mathcal{P}$ of $\mathfrak{S}_n$ with $P(\mathcal{C}) > 0$ for all $\mathcal{C}$:

$$W_{d_\tau}(P, P_{\mathcal{P}}) \leq$$

Distortion bound (Proposition 1)

Proposition 1

For any partition $\mathcal{P}$ of $\mathfrak{S}_n$ with $P(\mathcal{C}) > 0$ for all $\mathcal{C}$:

$$W_{d_\tau}(P,\, P_{\mathcal{P}}) \;\leq\; \mathcal{E}_P(\mathcal{P}) := \sum_{\mathcal{C}\in\mathcal{P}} P(\mathcal{C})\, V(\mathcal{C}) \;\leq\; 2\,\mathcal{E}'_P(\mathcal{P}) := 2\sum_{\mathcal{C}\in\mathcal{P}} P(\mathcal{C})\, V'(\mathcal{C}).$$

Proof sketch. Use the coupling $(\Sigma, \Sigma_{\mathcal{P}})$ with $\Sigma_{\mathcal{P}} = \sigma^*_{\mathcal{C}}$ whenever $\Sigma \in \mathcal{C}$. Then:

$$W_{d_\tau}(P, P_{\mathcal{P}}) \leq \mathbb{E}[d_\tau(\Sigma, \Sigma_{\mathcal{P}})] = \sum_{\mathcal{C}} P(\mathcal{C})\, \underbrace{\mathbb{E}[d_\tau(\Sigma, \sigma^*_{\mathcal{C}}) \mid \Sigma \in \mathcal{C}]}_{=V(\mathcal{C})}.$$

Three key properties of $\mathcal{E}'$:

- **Monotone:** $\mathcal{P}' \prec \mathcal{P} \;\Rightarrow\; \mathcal{E}'(\mathcal{P}') \leq \mathcal{E}'(\mathcal{P})$. Splitting never hurts.
- **Bounded:** $\mathcal{E}'(\mathcal{P}) \leq V'_P$ for any $\mathcal{P}$. Always better than the global median.
- **Computable:** depends only on local pairwise probabilities $p_{ij}(\mathcal{C})$.

Outline

1. Motivation
2. Background
3. Consensus Ranking Distributions
4. The k-medians algorithms
5. Experiments

The Exact k-medians Algorithm

Analogous to k-means but operating on $\mathfrak{S}_n$ with d_τ:

Algorithm (sketch)

1. **Initialization.** Sample K random local medians $\{\widehat{\sigma}_1^*, \ldots, \widehat{\sigma}_K^*\}$; form initial Voronoi partition $\mathcal{P}_0$.
2. **Iteration.** While $\mathcal{P}$ is not stable:
   (a) *Local medians.* For each cell $\mathcal{C}_\ell$, compute
   $$\widehat{\sigma}_\ell^* \in \operatorname*{arg\,min}_{\sigma \in \mathfrak{S}_n} \mathbb{E}_P[d(\Sigma, \sigma) \mid \Sigma \in \mathcal{C}_\ell] .$$
   (b) *Voronoi update.* Re-assign each σ to its nearest median.
3. **Output.** Stable partition $\mathcal{P}$ and medians $\{\widehat{\sigma}_1^*, \ldots, \widehat{\sigma}_K^*\}$.

Key challenge: Computing a Kemeny median per cluster is NP-hard in general.

Theoretical Guarantees

Proposition (Main result)

Let P be a distribution on $\mathfrak{S}_n$ and $(\mathcal{P}_p)_{p \geq 0}$ the sequence of partitions produced by the algorithm. Then:

1. **Finite convergence.** The algorithm terminates in a finite number of iterations.
2. **Wasserstein decrease.** $W_\tau(P, P_{\mathcal{P}_p})$ also decreases along iterations.
3. **Optimality.** Any minimizing partition of size K is a Voronoi partition w.r.t. its local medians.

Proof idea: Strict decrease of a finite-valued function over the finite set of K-partitions of $\mathrm{Supp}(P)$.

Outline

1. Motivation
2. Background
3. Consensus Ranking Distributions
4. The k-medians algorithms
5. Experiments

Numerical Experiments — Overview

Two algorithms tested:

- **Exact** k-medians
  (exact Kemeny median per cluster)
- **Approximate** k-medians
  (Borda rule surrogate)

Two data regimes:

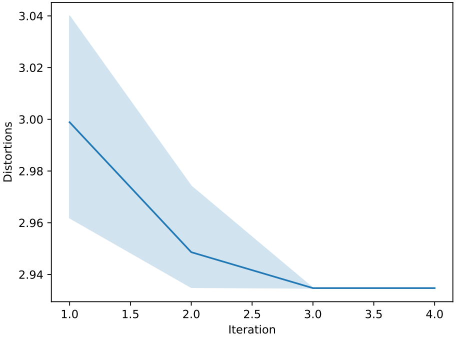
- **Synthetic:** mixtures of Mallows (MM) and Plackett-Luce (PL) with known ground-truth centers
- **Real:** Sushi preference dataset ($n = 10$, $\approx$5000 rankings)

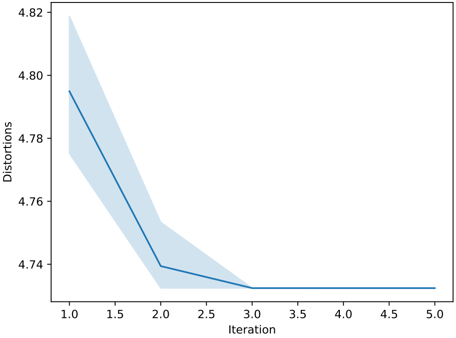
Metrics per iteration:

- Distortion $\mathcal{E}(\mathcal{P}_p)$
  (should decrease monotonically)
- Min. distance from estimated centers to true centers $\mathcal{C}$
  (should decrease)
- Co-membership matrix
  (should be block-diagonal)

Exact k-medians: Distortion & Distance to True Centers

Setup: Mixture of Mallows models, $\phi = 0.1$; exact local median per cluster; averaged over 10 runs.





Top: distortion. *Bottom:* distance to true centers. *Left:* $n = 6, K = 2$ ($\approx$30 s/run). *Right:* $n = 7, K = 3$ ($\approx$30 min/run).

Exact k-medians: Scalability Wall

Observation

Both distortion and distance to true centers decrease monotonically — confirming theory.
But: computation time grows factorially with n.

Setting	Time/iteration	Time/run
$n = 6,\ K = 2$	$\sim 7\,\mathrm{s}$	$\sim 30\,\mathrm{s}$
$n = 7,\ K = 3$	$\sim 7\,\mathrm{min}$	$\sim 30\,\mathrm{min}$

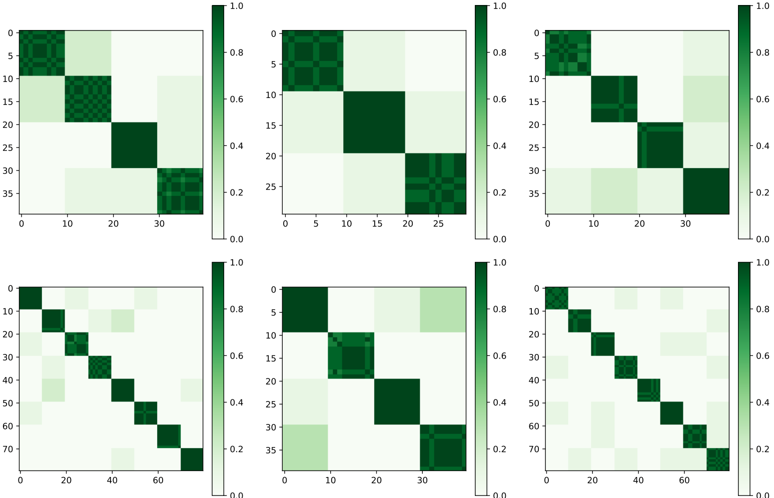
$\Rightarrow$ Need for approximation

The exact algorithm is impractical beyond $n \approx 7$. The approximate algorithm (Borda rule) handles $n \leq 100$ in seconds.

Approximate k-medians: Co-membership Matrix

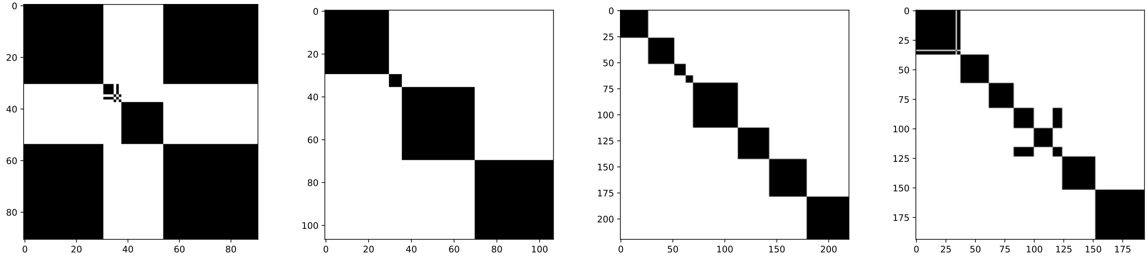
M_{ij} = fraction of runs where rankings i, j land in the same cluster.

Perfect recovery $\Leftrightarrow$ K-block-diagonal matrix.



Heterogeneous Components & Sensitivity to Initialization

Setup: Components differ in both *cardinality* and *dispersion* $\phi = 0.1$.
Two independent runs reported per setting (aggregation is not meaningful here).



(a,b) MM $n = 40$, $K = 4$. (c,d) MM $n = 100$, $K = 8$.

- Almost block-diagonal in most runs.
- Panel (a): two clusters partially merged — effect of initialization.
- **Different initializations recover the original partitions** (as in k-means).

Approximate k-medians: Computation Times

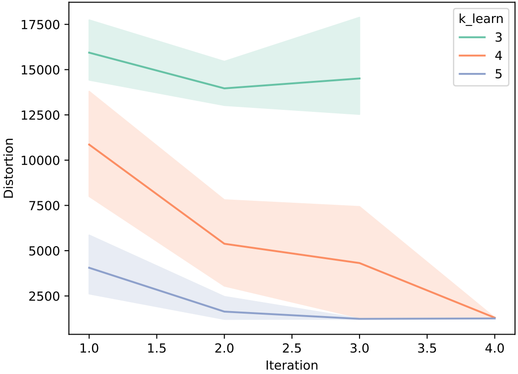
Speedup vs. exact

Exact: $\sim$30 min for $n = 7$, $K = 3$.
Approximate: < 8 s for $n = 100$, $K = 8$.

- Extends directly to **partial rankings** (distances and Borda medians both generalise).
- Complexity: $\mathcal{O}(n \log n \cdot N)$ per cluster, linear in sample size.

Choosing K: Distortion & Distance to Centers ($n = 100$)

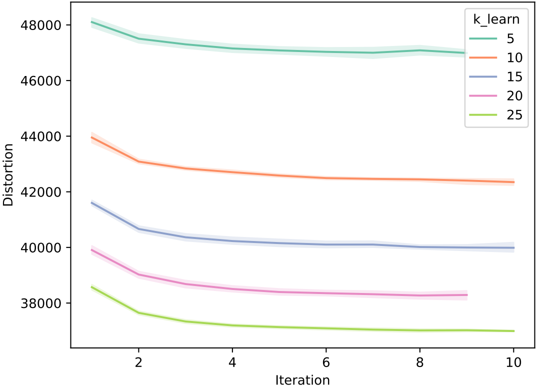
Setup: MM, $K_{\text{true}} = 4$, $n = 100$; algorithm run with $K \in \{3, 4, 5\}$, varying ϕ.



Columns: $\phi = 0.3$, 0.5, 0.7. *Top:* distortion. *Bottom:* distance to true centers.

Key takeaway: Convergence in few iterations for any K; diminishing returns as K grows.

Real Data: Sushi Preference Dataset

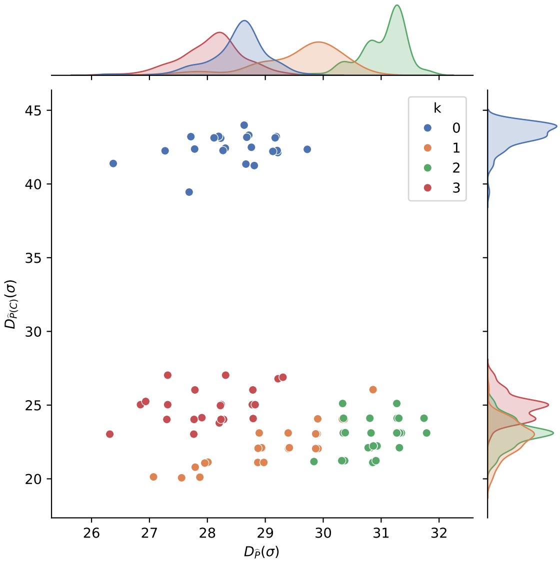


Dataset: Rankings of $n = 10$ sushi types by $\approx$5000 individuals.

Exact algorithm infeasible $\Rightarrow$ approximate algorithm, $K \in \{5, \ldots, 25\}$.

- Distortion converges in ≈ 10 iterations, **independent of K**.
- Larger K $\Rightarrow$ lower distortion, with **diminishing returns**.
- No sharp elbow: suggests a **diffuse** underlying preference distribution.

Depths


k
0
1
2
3
45
40
35
30
25
20
26
27
28
29
30
31
32
$D_{\hat{P}(C)}(\sigma)$
$D_{\hat{P}}(\sigma)$

Conclusion

We introduced a k-medians framework for ranking distributions, extending k-means to $(\mathfrak{S}_n, d_\tau)$:

- ✓ **No restrictive assumptions** on partition structure or component distributions.
- ✓ **Provable guarantees:** monotone distortion decrease, finite convergence, Voronoi optimality.
- ✓ **Scalable approximation:** Borda-based surrogate reduces complexity to $\mathcal{O}(n \log n \cdot N)$.
- ✓ **Empirically validated** on synthetic (MM, PL) and real (Sushi) datasets.

Future directions:

- Statistical convergence rates for the approximate algorithm.
- Principled model selection for K (penalization / information criteria).
- Extension to partial rankings and pairwise comparison data.
- Application to LLM benchmark aggregation and collective preference modeling.

Thank you!

ekhine.irurozki@telecom-paris.fr