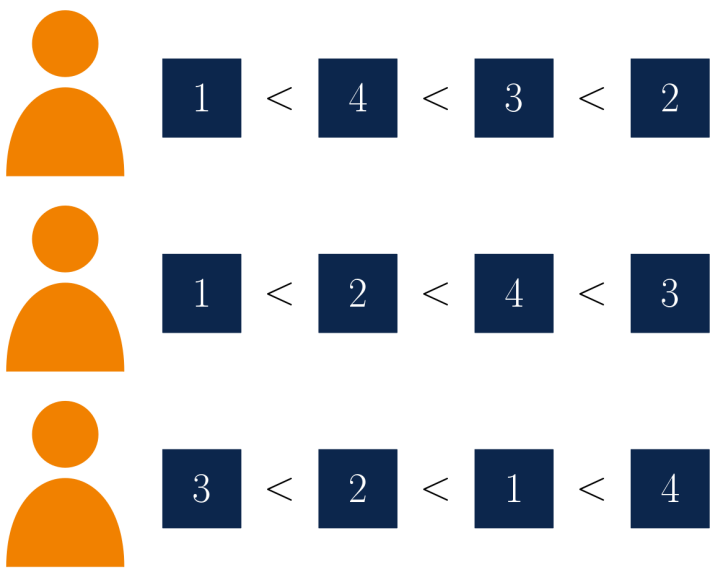
Aggregating Incomplete and Noisy Rankings

Dimitris Fotakis and Alkis Kalavasis and Konstantinos Stavropoulos

National Technical University of Athens

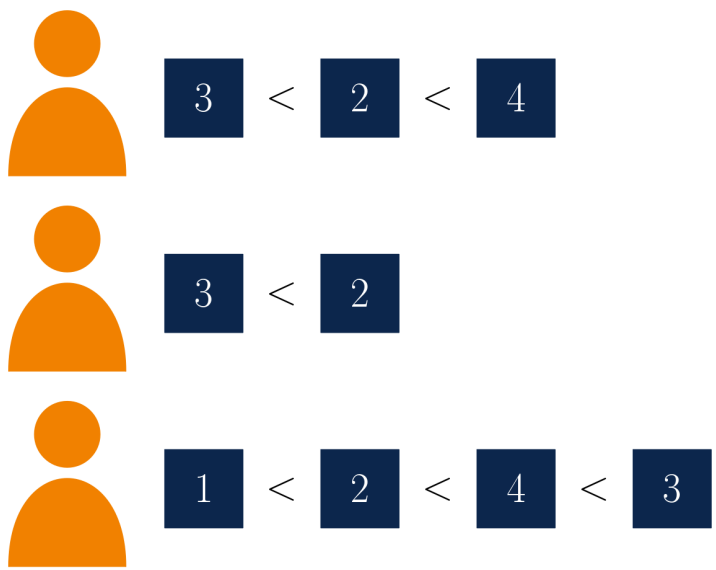
Problem

- Aggregating collections of ranked preferences: *analyze statistical ranking models* (e.g. [Mal57], [Thu27], [Smi50], [BT52], [Pla75], [Luc12]).



Problem

- Incomplete input rankings: *analyze incomplete ranking models.*



Mallows Model

- Extensively studied (e.g. [BM09], [AF98], [CPS13], [LM18], [BFZ19], [FV86], [MM03], [LL03]).
- Central (ground truth) ranking $\pi_0 \leftrightarrow$ ideal output of aggregation.
- Exponential decay to the (Kendall Tau) distance from π_0:

$$\Pr[\pi|\pi_0, \beta] \propto \exp(-\beta d_{KT}(\pi_0, \pi)), \forall \pi \in \mathfrak{S}_n, \tag{1}$$

$\beta > 0$: spread parameter,
$d_{KT}(\pi, \pi') = \#\text{inversions between } \pi, \pi'$.

Selective Mallows Model

- Generates incomplete rankings.
- Selection sets: $\mathcal{S} = (S_1, \ldots, S_r)$, where $S_i \subseteq [n]$.
- Exponential decay to the distance from sub-rankings of π_0:

$$\Pr[(\pi_1, \ldots, \pi_r) | \pi_0, \beta, \mathcal{S}] \propto \prod_{i \in [r]} \exp(-\beta d_{KT}(\pi_0|_{S_i}, \pi_i)), \qquad (2)$$

π_i: a ranking of the elements of S_i,
$\pi_0|_{S_i}$: the ranking of the elements of S_i according to π_0.

Results

- (Pair appearance) **frequency parameter** $p \in [0, 1]$:

$$|\{S_i : S_i \ni a, b\}| \geq p \cdot |\mathcal{S}|, \forall a, b \in [n]$$

Results: Sample Complexity

- For error prob. $\epsilon \in (0,1)$: $\text{poly}(1/\beta) \cdot \Theta(\frac{1}{p} \cdot \log(n/\epsilon))$.
- **Upper Bound:** $O(\frac{1}{p(1-e^{-\beta})^2} \cdot \log(n/\epsilon))$

  Large enough sample profile length

  $\Downarrow$

  Each pair correctly ranked by majority.
- **Lower Bound:** $\Omega(\frac{1}{\beta p} \cdot \log(n/\epsilon))$

  Small sample profile length

  $\Downarrow$

  $\exists$ rarely observed, large set of disjoint pairs.

Results: Maximum Likelihood Estimation

- As in (complete) Mallows: MLE of π_0 linked to an NP-hard problem.
- However:
  - Average case $\Rightarrow$ initial approximation of MLE.
  - Structure of problem $\Rightarrow$ efficient local search.
- Finally: For any $\alpha > 0$, we find MLE w.p. $\geq 1 - n^{-\alpha}$, in time:

$$T = O\left(n^2 + n^{1+O(\frac{2+\alpha}{r\beta p^4})} 2^{O(\frac{1}{p^3\beta})} \log^2 n\right)$$

Results: Top$-k$ Retrieval

- **Upper Bound:** $O\left(\frac{\log(k/\epsilon)}{p(1-e^{-\beta})^2} + \frac{\log(n/\epsilon)}{p^2\beta k}\right)$ (if $k = \omega(1/(p\beta))$)

  Low ranked alternatives ruled out with few samples.
- **Lower Bound:** $\Omega(\frac{1}{\beta p}\log(k/\epsilon))$

  Implied by sample complexity of finding π_0.

Positional Estimator

- Given a selective Mallows sample profile, PosEst:
  1. Calculates pairwise majority positions:

$$\hat{\pi}(a) = \{b \in [n] : b < a \text{ in most common appearances}\}$$

  2. Breaks ties uniformly.
- Optimal sample complexity for π_0, Upper bound for top$-k$ & Approximates π_0 (first step of MLE).

Experiments

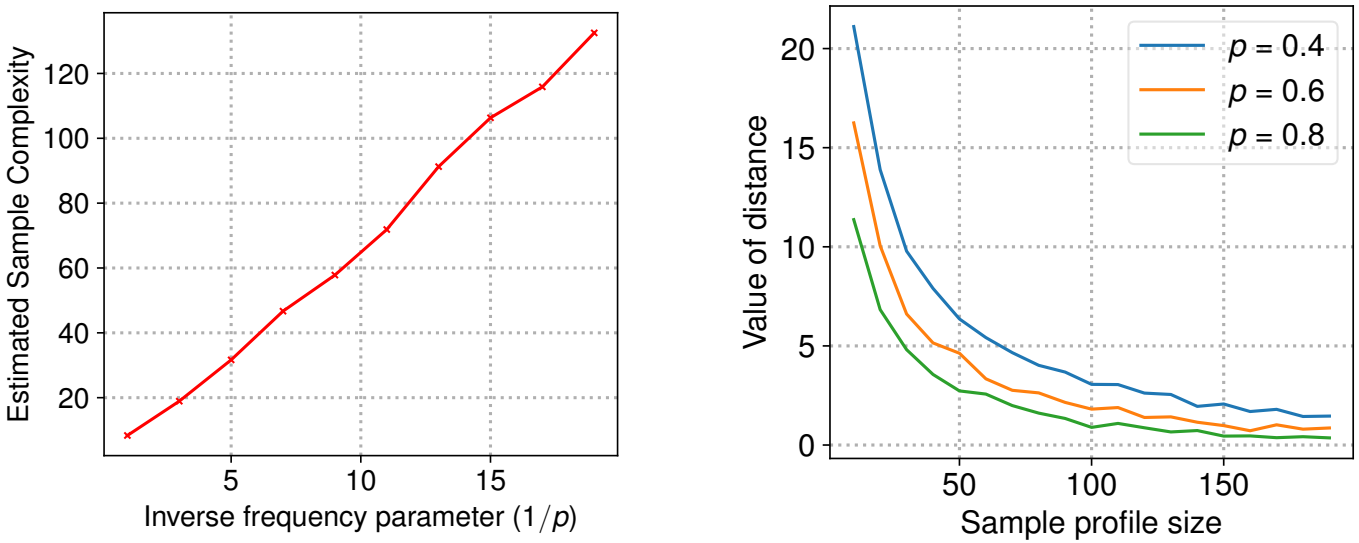


Figure: (Left:) Estimated sample complexity of retrieving, with probability at least 0.95 and using PosEst, the central ranking from selective Mallows samples, with $n = 20$, $\beta = 2$, over the frequency parameter's inverse. (Right:) Average Kendall Tau distance between the output of PosEst and the central ranking with respect to the size of the sample profile, for different values of the frequency parameter p, when $n = 20$, $\beta = 0.3$.

References

Colin L Mallows. (1957)
Non-Null Ranking Models. I.
Biometrika, 44(1/2):114—130, 1957.

Braverman, Mark and Mossel, Elchanan. (2009)
Sorting from Noisy Information
arXiv preprint arXiv:0910.1191, 2009.

Caragiannis, Ioannis and Procaccia, Ariel D and Shah, Nisarg. (2013)
When Do Noisy Votes Reveal the Truth?
Proceedings of the fourteenth ACM conference on Electronic commerce, 143–160, 2013.

Fligner, Michael A and Verducci, Joseph S. (1986)
Distance Based Ranking Models
Journal of the Royal Statistical Society: Series B (Methodological), 48(3):359–369, 1986.

References

Plackett, Robin L. (1975)

The Analysis of Permutations

Journal of the Royal Statistical Society: Series C (Applied Statistics), 24(2):193–202, 1975.

Murphy, Thomas Brendan and Martin, Donal. (2003)

Mixtures of distance-based models for ranking data

Computational statistics & data analysis, 41(3-4):645–655, 2003.

Smith, B Babington. (1950)

Discussion of Professor Ross's Paper

Journal of the Royal Statistical Society B, 12(1):41–59, 1950.

Bradley, Ralph Allan and Terry, Milton E. (1952)

Rank analysis of incomplete block designs: I. The method of paired comparisons

Biometrika, 39(3/4):324–345, 1952.

References

Braverman, Mark and Mossel, Elchanan. (2008)

Noisy Sorting Without Resampling

Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms, 268–276, 2008.

Busa-Fekete, Robert and Fotakis, Dimitris and Szörényi, Balázs and Zampetakis, Manolis. (2019)

Optimal Learning of Mallows Block Model

Conference on Learning Theory, 529–532, 2019.

Lebanon, Guy and Lafferty, John D. (2003)

Conditional Models on the Ranking Poset

Advances in Neural Information Processing Systems, 431–438, 2003.

Luce, R Duncan. (2012)

Individual Choice Behavior: A Theoretical Analysis

2012.

References

Thurstone, Louis L. (1927)

A Law of Comparative Judgment.

Psychological review, 34(4):273, 1927.

Adkins, Laura and Fligner, Michael. (1998)

A non-iterative procedure for maximum likelihood estimation of the parameters of Mallows' model based on partial rankings

Communications in Statistics-Theory and Methods, 27(9):2199–2220, 1998.

Liu, Allen and Moitra, Ankur. (2018)

Efficiently Learning Mixtures of Mallows Models

2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS), 627–638, 2018.

The End