Reconstructing Hidden Permutations Using the Average-Precision (AP) Correlation Statistic

Lorenzo De Stefani, Alessandro Epasto, Eli Upfal (Brown University), Fabio Vandin (University of Padova)
(lorenzo, epasto, eli)@cs.brown.edu, vadinfa@dei.unipd.it

**MOTIVATION**

Probabilistic models of rankings studied in:
- social sciences
- statistics
- machine learning
- computer science

Applications include:
- understanding user preferences
- ordering web search results
- aggregating crowd-sourcing data
- optimizing recommendation systems results

Most work in literature focused on the Mallows model.

**THE MALLOWS MODEL**

In the Mallows model $\mathcal{M}(\beta, \pi)$, the probability of observing a permutation $\sigma$ is inversely proportional to its tau distance from a central ("ground truth") permutation $\pi$.

The tau distance $d_k(\pi, \sigma)$ between permutations counts the number of items whose order is inverted.

All inversions are weighted in the same way!

In many cases, the order of items at the top of the ranking is more significant than the order of the items at the bottom!

**APPLICATION TO CLUSTERING**

We use the AP-model estimators for an unsupervised clustering algorithm based on k-means.
- We apply the algorithm to cluster web pages
- Purity and ROC converge to over 80% after a few iterations
- Our approach is resilient to noise in the lower position of the rankings

**APPLICATION TO CLASSIFICATION**

It has been observed that in the context of high-dimensional gene expressions data ($> 10^4$ genes), the relative order of the genes is more important than their absolute magnitude.

- We use the AP-model to classify gene expressions into one of two binary classes (e.g., "Normal vs. Tumor")
- AP distance improves over tau distance-based methods for most datasets!

---

**THE AP - MODEL**

We propose the new AP-model which uses the AP-statistics as measure of distance between permutations.

\[
\Pr_{\mathcal{M}(\beta, \pi)}(\sigma) = Z^{-1}_\beta \exp(-\beta d_{AP}(\pi, \sigma))
\]

The AP-statistics counts inversions weighting them according to the position of the swapped items in the central permutation $\pi$.

\[
d_{AP}(\pi, \sigma) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E_{ij} \frac{n}{2(j-1)}
\]

\[
d_k(\pi, \sigma) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E_{ij}
\]

---

**ALGORITHMS**

We provide efficient algorithms for reconstructing the central permutation $\pi$ on $n$ items from $O(\log_2 n)$ observations.

---

**Correlation Statistic**

- It has been observed that in the context of high-dimensional gene expressions data ($> 10^4$ genes), the relative order of the genes is more important than their absolute magnitude.
- We use the AP-model to classify gene expressions into one of two binary classes (e.g., "Normal vs. Tumor")
- AP distance improves over tau distance-based methods for most datasets!