Dimensionality Reduction and (Bucket) Ranking: A Mass Transportation Approach

Anna Korba [1,2] Mastane Achab [1] Stephan Clémençon [1]

[1] LTCI, Télécom ParisTech, Université Paris-Saclay

[2] Gatsby Unit, CSML, University College London

ALT 2019, Chicago

Outline

1. Introduction
2. Dimensionality Reduction on $\mathfrak{S}_n$
3. Theoretical results
4. Numerical Experiments on Real-world Preference Data

Outline

Introduction - Ranking Data

Consider a set of items $[\![n]\!] := \{1, \ldots, n\}$.

A ranking is an **ordered list** (of any size) **of items** of $[\![n]\!]$

 $\rightarrow$  $\prec$  $\prec$  $\prec$ 

Example: *travel* $\prec$ *sports* $\prec$ *finance* $\prec$ *clothing*

Introduction - Ranking Data

Consider a set of items $[\![n]\!] := \{1, \ldots, n\}$.

A ranking is an **ordered list** (of any size) **of items** of $[\![n]\!]$

 $\rightarrow$  $\prec$  $\prec$  $\prec$ 

Example: *travel* $\prec$ *sports* $\prec$ *finance* $\prec$ *clothing*

Many applications involve rankings/comparisons:

- Modelling human preferences (elections, surveys, online implicit feedback)
- Computer systems (search engines, recommendation systems)
- Other (competitions, biological data...)

Ranking data - Permutations

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

$$a_1 \prec a_2 \prec \cdots \prec a_n \qquad \Leftrightarrow \qquad \sigma \in \mathfrak{S}_n \text{ such that } \sigma(a_i) = i$$

$$2 \prec 1 \prec 3 \prec 4 \qquad \Leftrightarrow \qquad \sigma = 2134 \ \ (\sigma(2) = 1, \sigma(1) = 2, \dots)$$

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

Ranking data - Permutations

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

$$a_1 \prec a_2 \prec \dots \prec a_n \quad \Leftrightarrow \quad \sigma \in \mathfrak{S}_n \text{ such that } \sigma(a_i) = i$$

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

Let $\mathfrak{S}_n$ be set of permutations of $[\![n]\!]$, the symmetric group.

Ex: $\mathfrak{S}_4 = 1234, 1324, 1423, \dots, 4321$

$\Longrightarrow$ A distribution P on rankings/$\mathfrak{S}_n$ is described by an exploding number $(n! - 1)$ of parameters!

Ranking data - Permutations

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

$$a_1 \prec a_2 \prec \cdots \prec a_n \quad \Leftrightarrow \quad \sigma \in \mathfrak{S}_n \text{ such that } \sigma(a_i) = i$$

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

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

$\Longrightarrow$ A distribution P on rankings/$\mathfrak{S}_n$ is described by an exploding number $(n! - 1)$ of parameters!

How to summarize P ?

Dimensionality Reduction

- No vector space structure for permutations
- Dimensionality reduction methods usually rely on linear algebra (e.g. PCA)

Dimensionality Reduction

- No vector space structure for permutations
- Dimensionality reduction methods usually rely on linear algebra (e.g. PCA)

Our proposal

Summarize P on $\mathfrak{S}_n$ by:

- a bucket ordering (**a partial order**) $\mathcal{C}$
- a **sparse** ranking distribution $P_{\mathcal{C}}$

Outline

Background on Consensus Ranking

Dimensionality reduction techniques generally rest upon averages or linear combinations of the features, representing efficiently the data.

Background on Consensus Ranking

Dimensionality reduction techniques generally rest upon averages or linear combinations of the features, representing efficiently the data.

Find the dirac distribution closest to P:

$$\delta_{\sigma^*} = \min_{\sigma \in \mathfrak{S}_n} W_{d,q}\left(P, \delta_\sigma\right)$$

where $W_{d,q}\left(P, P'\right) = \inf_{\Sigma \sim P,\ \Sigma' \sim P'} \mathbb{E}\left[d^q(\Sigma, \Sigma')\right]$ is the Wassertein distance.

Background on Consensus Ranking

Dimensionality reduction techniques generally rest upon averages or linear combinations of the features, representing efficiently the data.

Find the dirac distribution closest to P:

$$\delta_{\sigma^*} = \min_{\sigma \in \mathfrak{S}_n} W_{d,q}\left(P, \delta_\sigma\right)$$

where $W_{d,q}\left(P, P'\right) = \inf_{\Sigma \sim P,\, \Sigma' \sim P'} \mathbb{E}\left[d^q(\Sigma, \Sigma')\right]$ is the Wassertein distance.

$\Longrightarrow W_{d,q}\left(P, \delta_\sigma\right) = \mathbb{E}_{\Sigma \sim P}[d(\Sigma, \sigma)]$.
$\Longrightarrow$ **ranking aggregation/consensus ranking** as a radical dimensionality reduction procedure

We choose the Kendall's τ distance:

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

Kemeny medians are solutions of:

$$\sigma^*_P = \min_{\sigma \in \mathfrak{S}_n} \sum_{1 \leq i < j \leq n} p_{i,j} \mathbb{I}\{\sigma(i) > \sigma(j)\} + (1 - p_{i,j}) \mathbb{I}\{\sigma(i) < \sigma(j)\} \quad (1)$$

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

We choose the Kendall's τ distance:

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

Kemeny medians are solutions of:

$$\sigma_P^* = \min_{\sigma \in \mathfrak{S}_n} \sum_{1 \leq i < j \leq n} p_{i,j} \mathbb{I}\{\sigma(i) > \sigma(j)\} + (1 - p_{i,j}) \mathbb{I}\{\sigma(i) < \sigma(j)\} \quad (1)$$

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

[Korba et al., 2017] $\Rightarrow$ (1) is given by Copeland ranking

$$\sigma_P^*(i) = 1 + \sum_{j \neq i} \mathbb{I}\{p_{i,j} < 1/2\}.$$

The rank of item i in σ_P^* is its number of pairwise defeats against other items

if P **strictly stochastically transitive**:

- $p_{i,j} \neq 1/2$ for all $i < j$
- $p_{i,j} \geq 1/2$ and $p_{j,k} \geq 1/2 \ \Rightarrow \ p_{i,k} \geq 1/2$

From ranking aggregation to bucket ranking

From ranking aggregation to bucket ranking

Let $\mathcal{C} = (\mathcal{C}_1, \ldots, \mathcal{C}_K)$ be a bucket order.

A bucket order $\mathcal{C} = (\mathcal{C}_1, \ldots, \mathcal{C}_K)$ is an ordered partition of $[\![n]\!]$:

- $\mathcal{C}_k$'s disjoint non empty subsets of $[\![n]\!]$
- $\cup_{k=1}^{K} \mathcal{C}_k = [\![n]\!]$

$\mathcal{C}$ is described by K (its *size*) and $(\#\mathcal{C}_1, \ldots, \#\mathcal{C}_K)$ (*shape*).

From ranking aggregation to bucket ranking

Let $\mathcal{C} = (\mathcal{C}_1, \ldots, \mathcal{C}_K)$ be a bucket order.

A bucket order $\mathcal{C} = (\mathcal{C}_1, \ldots, \mathcal{C}_K)$ is an ordered partition of $[\![n]\!]$:

- $\mathcal{C}_k$'s disjoint non empty subsets of $[\![n]\!]$
- $\cup_{k=1}^{K} \mathcal{C}_k = [\![n]\!]$

$\mathcal{C}$ is described by K (its *size*) and $(\#\mathcal{C}_1, \ldots, \#\mathcal{C}_K)$ (*shape*).

Find the distribution $P_{\mathcal{C}}$ closest to P:

$$\Lambda_P(\mathcal{C}) = \min_{P' \in \mathbf{P}_{\mathcal{C}}} W_{d_\tau,1}(P, P')$$

where $\mathbf{P}_{\mathcal{C}}$ set of distributions associated to $\mathcal{C}$.

Sparsity and Bucket orders

Sparse distributions

$\mathbf{P}_{\mathcal{C}}$: set of all bucket distributions P' associated to $\mathcal{C}$

- P' distribution on $\mathfrak{S}_n$
- if $i \prec_{\mathcal{C}} j$ (i.e. $\exists k < l$, s.t. $(i,j) \in (\mathcal{C}_k, \mathcal{C}_l)$), then $p'_{j,i} = \mathbb{P}_{\Sigma' \sim P'}\{\Sigma'(j) < \Sigma'(i)\} = 0$

i.e. the order of two items in two $\neq$ buckets is deterministic

$\Longrightarrow P' \in \mathbf{P}_{\mathcal{C}}$ described by $d_{\mathcal{C}} = \prod_{1 \leq k \leq K} \#\mathcal{C}_k! - 1 \leq n! - 1$ parameters

Dimensionality reduction with optimal coupling

Proposition (Optimal Coupling)

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

optimal when $P' = P_{\mathcal{C}}$ the distribution of $\Sigma_{\mathcal{C}}$:

$\forall k \in \{1, \ldots, K\}, \forall i \in \mathcal{C}_k, \;\; \Sigma_{\mathcal{C}}(i) = 1 + \sum_{l<k} \#\mathcal{C}_l + \sum_{j \in \mathcal{C}_k} \mathbb{I}\{\Sigma(j) < \Sigma(i)\},$

Dimensionality reduction with optimal coupling

Proposition (Optimal Coupling)

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

optimal when $P' = P_{\mathcal{C}}$ the distribution of $\Sigma_{\mathcal{C}}$:

$\forall k \in \{1, \ldots, K\}, \forall i \in \mathcal{C}_k, \;\; \Sigma_{\mathcal{C}}(i) = 1 + \sum_{l<k} \#\mathcal{C}_l + \sum_{j \in \mathcal{C}_k} \mathbb{I}\{\Sigma(j) < \Sigma(i)\},$

Dimensionality Reduction

Let $K \leq n$ and $\mathbf{C}_{K,\lambda}$ the set of all bucket orders of size K and shape λ. A natural dimensionality reduction approach consists in finding a solution $C^{*(K)}$ of:

$$\min_{\mathcal{C} \in \mathbf{C}_{K,\lambda}} \Lambda_P(\mathcal{C})$$

as well as a solution $P_{C^{*(K)}}$ of $\Lambda_P(C^{*(K)})$ and a coupling $(\Sigma, \Sigma_{C^{*(K)}})$ s.t. $\mathbb{E}\left[d_\tau(\Sigma, \Sigma_{C^{*(K)}})\right]$.

Outline

Optimality

Assume that P is strongly (and strictly*) stochastically transitive i.e.:

$$p_{i,j} \geq 1/2 \text{ and } p_{j,k} \geq 1/2 \;\Rightarrow\; p_{i,k} \geq max(p_{i,j}, p_{j,k}).$$

*: $p_{i,j} \neq 1/2$.

Optimality

Assume that P is strongly (and strictly*) stochastically transitive i.e.:

$$p_{i,j} \geq 1/2 \text{ and } p_{j,k} \geq 1/2 \Rightarrow p_{i,k} \geq max(p_{i,j}, p_{j,k}).$$

*: $p_{i,j} \neq 1/2$.

Theorem

(i) $\Lambda_P(\cdot)$ *has a unique minimizer* $\mathcal{C}^{*(K,\lambda)}$ *over* $\mathbf{C}_{K,\lambda}$.

(ii) $\mathcal{C}^{*(K,\lambda)}$ *is the unique bucket order in* $\mathbf{C}_{K,\lambda}$ *agreeing with the Kemeny median* σ^*_P*:* $\mathcal{C}^{*(K,\lambda)} = (\mathcal{C}_1^{*(K,\lambda)}, \ldots, \mathcal{C}_K^{*(K,\lambda)})$*, where*

$$\mathcal{C}_k^{*(K,\lambda)} = \left\{ i \in [\![n]\!] : \sum_{l<k} \lambda_l < \sigma^*_P(i) \leq \sum_{l \leq k} \lambda_l \right\} \text{ for } k \in \{1, \ldots, K\}.$$

Optimality

Assume that P is strongly (and strictly*) stochastically transitive i.e.:

$$p_{i,j} \geq 1/2 \text{ and } p_{j,k} \geq 1/2 \;\Rightarrow\; p_{i,k} \geq max(p_{i,j}, p_{j,k}).$$

*: $p_{i,j} \neq 1/2$.

Theorem

(i) *$\Lambda_P(\cdot)$ has a unique minimizer $\mathcal{C}^{*(K,\lambda)}$ over $\mathbf{C}_{K,\lambda}$.*

(ii) *$\mathcal{C}^{*(K,\lambda)}$ is the unique bucket order in $\mathbf{C}_{K,\lambda}$ agreeing with the Kemeny median σ_P^*: $\mathcal{C}^{*(K,\lambda)} = (\mathcal{C}_1^{*(K,\lambda)}, \ \dots, \ \mathcal{C}_K^{*(K,\lambda)})$, where*

$$\mathcal{C}_k^{*(K,\lambda)} = \left\{ i \in [\![n]\!] : \sum_{l<k} \lambda_l < \sigma_P^*(i) \leq \sum_{l \leq k} \lambda_l \right\} \text{ for } k \in \{1, \dots, K\}.$$

$\Longrightarrow$ this result will lead to our practical method

Empirical setting

How to recover optimal buckets from a training sample $\Sigma_1, \dots, \Sigma_N \sim P$?

- Empirical pairwise probabilities:

$$\widehat{p}_{i,j} = \frac{1}{N} \sum_{s=1}^{N} \mathbb{I}\{\Sigma_s(i) < \Sigma_s(j)\}.$$

- Empirical distortion of any bucket order $\mathcal{C}$:

$$\widehat{\Lambda}_N(\mathcal{C}) = \Lambda_{\widehat{P}_N}(\mathcal{C}) = \sum_{1 \le k < l \le K} \sum_{(i,j) \in \mathcal{C}_k \times \mathcal{C}_l} \widehat{p}_{j,i}.$$

Rate bound

Empirical distortion minimizer $\widehat{C}_{K,\lambda}$ is solution of:

$$\min_{\mathcal{C} \in \mathbf{C}_{K,\lambda}} \widehat{\Lambda}_N(\mathcal{C}),$$

where $\mathbf{C}_{K,\lambda}$ set of bucket orders $\mathcal{C}$ of size K and shape $\lambda = (\lambda_1, \ldots, \lambda_K)$ (i.e. $\#\mathcal{C}_k = \lambda_k$ for all $1 \leq k \leq K$).

Rate bound

Empirical distortion minimizer $\widehat{C}_{K,\lambda}$ is solution of:

$$\min_{\mathcal{C} \in \mathbf{C}_{K,\lambda}} \widehat{\Lambda}_N(\mathcal{C}),$$

where $\mathbf{C}_{K,\lambda}$ set of bucket orders $\mathcal{C}$ of size K and shape $\lambda = (\lambda_1, \ldots, \lambda_K)$ (i.e. $\#\mathcal{C}_k = \lambda_k$ for all $1 \leq k \leq K$).

Theorem

For all $\delta \in (0,1)$, we have with probability at least $1-\delta$:

$$\Lambda_P(\widehat{C}_{K,\lambda}) - \inf_{\mathcal{C} \in \mathbf{C}_{K,\lambda}} \Lambda_P(\mathcal{C}) \leq \beta(n,\lambda) \times \sqrt{\frac{\log(\frac{1}{\delta})}{N}}.$$

Outline

Experiments

Sushi dataset (Kamishima, 2003):

- $n = 10$ sushi dishes
- $N = 5000$ full rankings.

Cars dataset

- $n = 10$ cars
- $N = 2500$ pairwise comparisons.

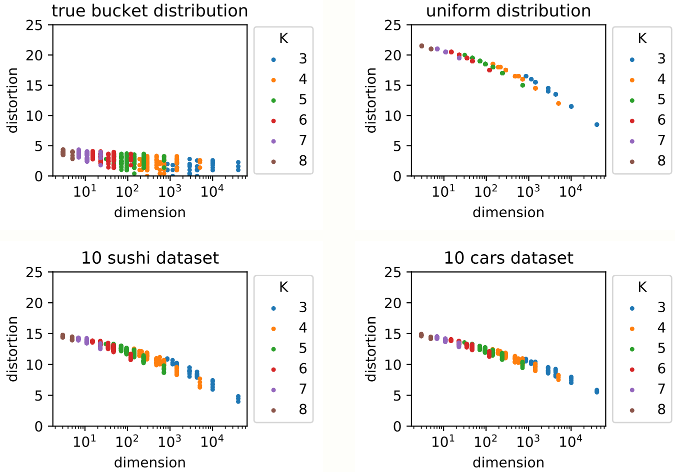
Method

1. Compute empirical pairwise probabilities $\widehat{p}_{i,j}$
2. Compute $\sigma_{\widehat{P}_N}$ with Copeland method

$$\sigma^*_{\widehat{P}_N}(i) = 1 + \sum_{j \neq i} \mathbb{I}\{\widehat{p}_{i,j} < 1/2\}.$$

3. Choose a size K, shape λ and segment $\sigma_{\widehat{P}_N}$ according to λ

Dimension-Distortion plot - $n = 10$ items



On top: true bucket distribution and uniform distribution.
Below: real preference data.

Conclusion

This paper introduces:

- theoretical concepts to represent in a sparse manner ranking distributions (bucket distributions)
- a distortion measure based on a mass transportation metric (Wassertein), to evaluate the accuracy of these representations

Future work: investigate how to exploit such representations in some tasks (e.g. clustering, ranking prediction)

Thank you!

Korba, A., Clémençon, S., and Sibony, E. (2017).
A learning theory of ranking aggregation.
In *Artificial Intelligence and Statistics*, pages 1001–1010.