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

Mastane Achab, Anna Korba, Stephan Clémençon

Outline

Outline

Introduction (1/2)

- Permutations over n items $[\![n]\!] = \{1, \ldots, n\}$

Introduction (1/2)

- Permutations over n items $[\![n]\!] = \{1, \ldots, n\}$
- Number of permutations explodes: $\#\mathfrak{S}_n = n!$

Introduction (1/2)

- Permutations over n items $[\![n]\!] = \{1, \dots, n\}$
- Number of permutations explodes: $\#\mathfrak{S}_n = n!$
- Distribution P on $\mathfrak{S}_n$: $n! - 1$ parameters

Introduction (2/2)

- Question: "How to summarize P?"

Introduction (2/2)

- Question: "How to summarize P?"
- Answer: dimensionality reduction

Introduction (2/2)

- Question: "How to summarize P?"
- Answer: dimensionality reduction
- Problem: no vector space structure for permutations

Outline

Preliminaries (1/2)

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

Preliminaries (1/2)

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

- $\mathcal{C}_i$'s disjoint non empty subsets of $[\![n]\!]$

Preliminaries (1/2)

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

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

Preliminaries (1/2)

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

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

Preliminaries (1/2)

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

- $\mathcal{C}_i$'s disjoint non empty subsets of $[\![n]\!]$
- $\cup_{k=1}^{K} \mathcal{C}_k = [\![n]\!]$
- K: "size" of $\mathcal{C}$
- $(\#\mathcal{C}_1, \ldots, \#\mathcal{C}_K)$: "shape" of $\mathcal{C}$

Preliminaries (1/2)

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

- $\mathcal{C}_i$'s disjoint non empty subsets of $[\![n]\!]$
- $\cup_{k=1}^{K} \mathcal{C}_k = [\![n]\!]$
- K: "size" of $\mathcal{C}$
- $(\#\mathcal{C}_1, \ldots, \#\mathcal{C}_K)$: "shape" of $\mathcal{C}$

Partial order: "i is ranked lower than j in $\mathcal{C}$" if $\exists k < l$ s.t. $(i, j) \in \mathcal{C}_k \times \mathcal{C}_l$.

Preliminaries (2/2)

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

Preliminaries (2/2)

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

- P' distribution on $\mathfrak{S}_n$

Preliminaries (2/2)

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

- P' distribution on $\mathfrak{S}_n$
- if $(i, j) \in \mathcal{C}_k \times \mathcal{C}_l$ $(k < l)$, then $p'_{j,i} = 0$

Preliminaries (2/2)

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

- P' distribution on $\mathfrak{S}_n$
- if $(i, j) \in \mathcal{C}_k \times \mathcal{C}_l$ $(k < l)$, then $p'_{j,i} = 0$
- $p'_{i,j} = \mathbb{P}(\Sigma'(i) < \Sigma'(j))$ for $\Sigma' \sim P'$

Preliminaries (2/2)

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

- P' distribution on $\mathfrak{S}_n$
- if $(i,j) \in \mathcal{C}_k \times \mathcal{C}_l$ $(k < l)$, then $p'_{j,i} = 0$
- $p'_{i,j} = \mathbb{P}(\Sigma'(i) < \Sigma'(j))$ for $\Sigma' \sim P'$

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

Background on Consensus Ranking

Consensus ranking (or "ranking aggregation"): summarize permutations $\sigma_1, \ldots, \sigma_N$ by a consensus/median ranking $\sigma^* \in \mathfrak{S}_n$ by solving:

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

Background on Consensus Ranking

Consensus ranking (or "ranking aggregation"): summarize permutations $\sigma_1, \ldots, \sigma_N$ by a consensus/median ranking $\sigma^* \in \mathfrak{S}_n$ by solving:

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

If $\Sigma_1, \ldots, \Sigma_N$ i.i.d. sampled from P (Korba et al., 2017), solve:

$$\min_{\sigma \in \mathfrak{S}_n} \mathbb{E}_{\Sigma \sim P}[d(\Sigma, \sigma)].$$

Kemeny medians

Particular choice for metric d:

Kemeny medians

Particular choice for metric d:

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

Kemeny medians

Particular choice for metric d:

- 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: $\min_{\sigma \in \mathfrak{S}_n} \mathbb{E}_{\Sigma \sim P}[d_\tau(\Sigma, \sigma)]$.

Kemeny medians

Particular choice for metric d:

- 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: $\min_{\sigma \in \mathfrak{S}_n} \mathbb{E}_{\Sigma \sim P}[d_\tau(\Sigma, \sigma)]$.

Unique Kemeny median σ^*_P if P strictly stochastically transitive:

Kemeny medians

Particular choice for metric d:

- 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: $\min_{\sigma \in \mathfrak{S}_n} \mathbb{E}_{\Sigma \sim P}[d_\tau(\Sigma, \sigma)]$.

Unique Kemeny median σ^*_P if P strictly stochastically transitive:

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

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

Bucket orders of size n

Consensus ranking: extreme case of bucket order $\mathcal{C}$ of size n.

Bucket orders of size n

Consensus ranking: extreme case of bucket order $\mathcal{C}$ of size n.

- $\mathcal{C} = (\{\sigma^{*-1}(1)\}, \ldots, \{\sigma^{*-1}(n)\})$

Bucket orders of size n

Consensus ranking: extreme case of bucket order $\mathcal{C}$ of size n.

- $\mathcal{C} = (\{\sigma^{*-1}(1)\}, \ldots, \{\sigma^{*-1}(n)\})$
- $\mathbf{P}_{\mathcal{C}} = \{\delta_{\sigma^*}\}$, hence dimension $d_{\mathcal{C}} = 0$

Bucket orders of size n

Consensus ranking: extreme case of bucket order $\mathcal{C}$ of size n.

- $\mathcal{C} = (\{\sigma^{*-1}(1)\}, \dots, \{\sigma^{*-1}(n)\})$
- $\mathbf{P}_{\mathcal{C}} = \{\delta_{\sigma^*}\}$, hence dimension $d_{\mathcal{C}} = 0$

Problem: generalization for any bucket order.

A Mass Transportation Approach

- Question: "How to quantify approximation error between orginal distrib. P and bucket distrib. $P' \in \mathbf{P}_{\mathcal{C}}$?".

A Mass Transportation Approach

- Question: "How to quantify approximation error between orginal distrib. P and bucket distrib. $P' \in \mathbf{P}_{\mathcal{C}}$?".
- Our answer: Wasserstein distance $W_{d,q}\left(P, P'\right)$.

Definition

$$W_{d,q}\left(P, P'\right) = \inf_{\Sigma \sim P,\, \Sigma' \sim P'} \mathbb{E}\left[d^q(\Sigma, \Sigma')\right]$$

A Mass Transportation Approach

- Question: "How to quantify approximation error between orginal distrib. P and bucket distrib. $P' \in \mathbf{P}_{\mathcal{C}}$?".
- Our answer: Wasserstein distance $W_{d,q}\left(P, P'\right)$.

Definition

$$W_{d,q}\left(P, P'\right) = \inf_{\Sigma \sim P,\, \Sigma' \sim P'} \mathbb{E}\left[d^q(\Sigma, \Sigma')\right]$$

- Why: because it generalizes consensus ranking. Indeed:

$$W_{d,1}\left(P, \delta_\sigma\right) = \mathbb{E}_{\Sigma \sim P}[d(\Sigma, \sigma)].$$

A Mass Transportation Approach

- Question: "How to quantify approximation error between orginal distrib. P and bucket distrib. $P' \in \mathbf{P}_{\mathcal{C}}$?".
- Our answer: Wasserstein distance $W_{d,q}(P, P')$.

Definition

$$W_{d,q}(P, P') = \inf_{\Sigma \sim P,\ \Sigma' \sim P'} \mathbb{E}\left[d^q(\Sigma, \Sigma')\right]$$

- Why: because it generalizes consensus ranking. Indeed:

$$W_{d,1}(P, \delta_\sigma) = \mathbb{E}_{\Sigma \sim P}[d(\Sigma, \sigma)].$$

- Focus on $d = d_\tau$ and $q = 1$.

Distortion measure

A bucket order $\mathcal{C}$ represents well P if small distortion $\Lambda_P(\mathcal{C})$.

Definition

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

Distortion measure

A bucket order $\mathcal{C}$ represents well P if small distortion $\Lambda_P(\mathcal{C})$.

Definition

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

Explicit expression for $\Lambda_P(\mathcal{C})$:

Proposition

$$\Lambda_P(\mathcal{C}) = \sum_{1 \leq k < l \leq K} \sum_{(i,j) \in \mathcal{C}_k \times \mathcal{C}_l} p_{j,i}$$

.

Outline

Empirical setting

Training sample: $\Sigma_1, \ldots, \Sigma_N$ i.i.d. from P.

Empirical setting

Training sample: $\Sigma_1, \ldots, \Sigma_N$ i.i.d. from P.

- Empirical pairwise probabilities:

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

Empirical setting

Training sample: $\Sigma_1, \ldots, \Sigma_N$ i.i.d. from 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 $\mathcal{C}$:

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

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 λ (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 λ (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 $1-\delta$:

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

The Strong Stochastic Transitive Case

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

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

The Strong Stochastic Transitive Case

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

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

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.*

The Strong Stochastic Transitive Case

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

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

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.*

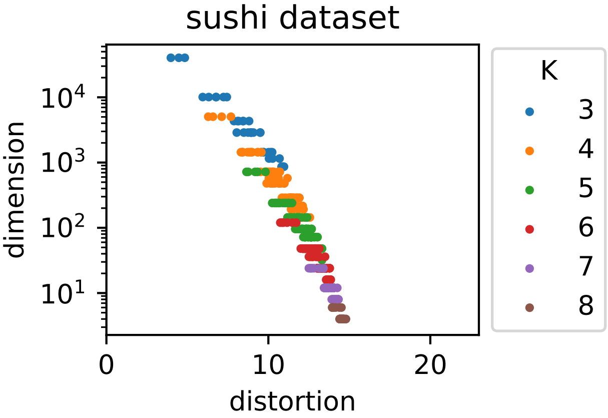
Consequence: agglomerative algorithm.

Outline

Experiments

Sushi dataset (Kamishima, 2003):

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



Thank you!