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

Mastane Achab, Anna Korba, Stephan Clémençon
LTCI, Télécom ParisTech
75013, Paris, France
first.last@telecom-paristech.fr

Abstract. Whereas most dimensionality reduction techniques (*e.g.* PCA) for multivariate data essentially rely on linear algebra to a certain extent, summarizing ranking data, viewed as realizations of a random permutation Σ on a set of items indexed by $i \in \{1, \ldots, n\}$, is a great statistical challenge, due to the absence of vector space structure for the set of permutations $\mathfrak{S}_n$. It is the goal of this article to develop an original framework for possibly reducing the number of parameters required to describe the distribution of a statistical population composed of rankings/permutations, on the premise that the collection of items under study can be partitioned into subsets/buckets, such that, with high probability, items in a certain bucket are either all ranked higher or else all ranked lower than items in another bucket. In this context, Σ's distribution can be hopefully represented in a sparse manner by a *bucket distribution*, *i.e.* a bucket ordering plus the ranking distributions within each bucket. More precisely, we introduce a dedicated distortion measure, based on a mass transportation metric, in order to quantify the accuracy of such representations.

1 Introduction

Many problems in machine learning and statistics naturally involve ranking data, expressing *preferences* over a set of items: elections, recommendation systems, search engines. In the case of full rankings (i.e. all the items are involved in the rankings), these data can be seen as permutations over the set of items. Because the number of possible rankings explodes with the number of items, it is of crucial importance to elaborate dedicated dimensionality reduction methods in order to represent ranking data efficiently. Indeed, it is far from straightforward to adapt traditional techniques such as Principal Component Analysis and its numerous variants to the ranking setup, the main barrier being the absence of a vector space structure on the set of permutations. In this paper, we develop a novel framework for representing the distribution of ranking data in a simple manner, that is shown to extend, in some sense, consensus ranking. The rationale behind the approach we promote is that, in many situations encountered in practice, the set of items may be partitioned into subsets/buckets, such that, with high probability, items belonging to a certain bucket are either all ranked higher or else all ranked lower than items lying in another bucket. In such a case, the ranking distribution can be described in a sparse fashion by: 1) a partial ordering structure (related to the buckets) and 2) the marginal ranking distributions associated to each bucket. Precisely, optimal representations are defined here as those associated to a bucket order minimizing a certain distortion measure we introduce, the latter being based on a mass transportation metric on the set of ranking distributions. In this paper, we also establish rate bounds describing the generalization capacity of bucket order representations obtained by minimizing an empirical version of the distortion and address model selection issues related to the choice of the bucket order size/shape. Numerical results are also displayed, providing in particular strong empirical evidence of the relevance of the notion of sparsity considered, which the dimensionality reduction technique introduced is based on.

The article is organized as follows. In section 2, a few concepts on ranking data are briefly recalled and the extended framework we consider for dimensionality reduction in the ranking context is described. In section 3 we give an insight about the results we obtained in this context: statistical results guaranteeing that optimal representations of reduced dimension can be learnt from ranking observations, and numerical experiments for illustration purpose.

2 Dimensionality Reduction on $\mathfrak{S}_n$

2.1 Preliminaries

It is the purpose of this section to introduce the main concepts and definitions that shall be used in the subsequent analysis. The indicator function of any event $\mathcal{E}$ is denoted by $\mathbb{I}\{\mathcal{E}\}$, the Dirac mass at any point a by δ_a, the cardinality of any finite subset A by $\#A$. Here and throughout, a full ranking on a set of items indexed by $[\![n]\!] = \{1, \ldots, n\}$ is seen as the permutation $\sigma \in \mathfrak{S}_n$ that maps any item i to its rank $\sigma(i)$. For any non empty subset $\mathcal{I} \subset [\![n]\!]$, any ranking σ on $[\![n]\!]$ naturally defines a ranking on $\mathcal{I}$, denoted by $\Pi_{\mathcal{I}}(\sigma)$ (*i.e.* $\forall i \in \mathcal{I}$, $\Pi_{\mathcal{I}}(\sigma)(i) = 1 + \sum_{j \in \mathcal{I} \setminus \{i\}} \mathbb{I}\{\sigma(j) < \sigma(i)\}$). If Σ is a random permutation on $\mathfrak{S}_n$ with distribution P, the distribution of $\Pi_{\mathcal{I}}(\Sigma)$ will be referred to as the marginal of P related to the subset $\mathcal{I}$. In particular, over a pair of items $(i, j) \in [\![n]\!]$, one can define the pairwise probability that item i is preferred to (i.e. ranked lower) item j: $p_{i,j} = \mathbb{P}\{\Sigma(i) < \Sigma(j)\}$ (so $p_{i,j} + p_{j,i} = 1$).

A bucket order $\mathcal{C}$ (also referred as a partial ranking in the literature) is a strict partial order defined by an ordered partition of $[\![n]\!]$, *i.e* a sequence $(\mathcal{C}_1, \ldots, \mathcal{C}_K)$ of $K \geq 1$ pairwise disjoint non empty subsets (buckets) of $[\![n]\!]$ such that: (1) $\cup_{k=1}^{K} \mathcal{C}_k = [\![n]\!]$, (2) $\forall (i,j) \in [\![n]\!]^2$, we have: $i \prec_{\mathcal{C}} j$ (i is ranked lower than j in $\mathcal{C}$) iff $\exists k < l$ s.t. $(i,j) \in \mathcal{C}_k \times \mathcal{C}_l$. The items in $\mathcal{C}_1$ thus have the lowest ranks whereas the items in $\mathcal{C}_K$ have the highest ones; and the items within each bucket are incomparable. For any bucket order $\mathcal{C} = (\mathcal{C}_1, \ldots, \mathcal{C}_K)$, its number of buckets K is referred to as its *size*, whereas the vector $\lambda = (\#\mathcal{C}_1, \ldots, \#\mathcal{C}_K)$, i.e the sequence of sizes of buckets in $\mathcal{C}$ (verifying $\sum_{k=1}^{K} \#\mathcal{C}_k = n$), is referred to as its *shape*. Hence, any bucket order $\mathcal{C}$ of size n corresponds to a full ranking/permutation $\sigma \in \mathfrak{S}_n$, whereas the set of all items $[\![n]\!]$ is the unique bucket order of size 1.

2.2 A Mass Transportation Approach

We now develop a framework for *dimensionality reduction* fully tailored to ranking data exhibiting a specific type of *sparsity*. For this purpose, we consider the so-termed *mass transportation* approach to defining metrics on the set of probability distributions on $\mathfrak{S}_n$ as follows (see (3)).

Definition 1 *Let $d : \mathfrak{S}_n^2 \to \mathbb{R}_+$ be a metric on $\mathfrak{S}_n$ and $q \geq 1$. The q-th Wasserstein metric with d as cost function between two probability distributions P and P' on $\mathfrak{S}_n$ is given by:*

$$W_{d,q}\left(P, P'\right) = \inf_{\Sigma \sim P,\ \Sigma' \sim P'} \mathbb{E}\left[d^q(\Sigma, \Sigma')\right], \tag{1}$$

where the infimum is taken over all possible couplings[1] *(Σ, Σ') of (P, P').*

As revealed by the following result, when the cost function d is equal to the Kendall's τ distance d_τ (see (1)) defined by:

$$\forall (\sigma, \sigma') \in \mathfrak{S}_n^2, \quad d_\tau(\sigma, \sigma') = \sum_{i<j} \mathbb{I}\{(\sigma(i) - \sigma(j))(\sigma'(i) - \sigma'(j)) < 0\},$$

which case the subsequent analysis focuses on, the Wasserstein metric is bounded by below by the l_1 distance between the pairwise probabilities.

Lemma 1 *For any probability distributions P and P' on $\mathfrak{S}_n$:*

$$W_{d_\tau,1}\left(P, P'\right) \geq \sum_{i<j} |p_{i,j} - p'_{i,j}|. \tag{2}$$

The equality holds true when the distribution P' is deterministic (i.e. when $\exists \sigma \in \mathfrak{S}_n$ s.t. $P' = \delta_\sigma$).

Sparsity and Bucket Orders. Here, we propose a way of describing a distribution P on $\mathfrak{S}_n$, originally described by $n! - 1$ parameters, by finding a much simpler distribution that approximates P in the sense of the Wasserstein metric introduced above under specific assumptions, extending somehow the consensus ranking concept. Let $K \leq n$ and $\mathcal{C} = (\mathcal{C}_1, \ldots, \mathcal{C}_K)$ be a *bucket order* of $[\![n]\!]$ with K buckets. In order to gain insight into the rationale behind the approach we promote, observe that, when $K \geq 2$, a distribution P' can be naturally said to be *sparse* if, for all $1 \leq k < l \leq K$ and all $(i,j) \in \mathcal{C}_k \times \mathcal{C}_l$, we have with probability one: $\Sigma'(i) < \Sigma'(j)$, when $\Sigma' \sim P'$ (or, equivalently, the probability that j is ranked lower that i is $p'_{j,i} = 0$). This means that the relative order of two items belonging to two different buckets is deterministic. Throughout the paper, such a probability distribution is referred to as a *bucket distribution* associated to $\mathcal{C}$. Since the variability of a bucket distribution corresponds to the variability of its marginals Π within each bucket, the set $\mathbf{P}_\mathcal{C}$ of all bucket distributions associated to $\mathcal{C}$ is of dimension $d_\mathcal{C} = \prod_{k \leq K} \#\mathcal{C}_k! - 1 \leq n! - 1$. A best summary in $\mathbf{P}_\mathcal{C}$ of a distribution P on $\mathfrak{S}_n$, in the sense of the Wasserstein metric (Eq. (1)), is then given by any solution $P^*_\mathcal{C}$ of the minimization problem

$$\min_{P' \in \mathbf{P}_\mathcal{C}} W_{d_\tau,1}(P, P'). \tag{3}$$

Set the distortion $\Lambda_P(\mathcal{C}) = \min_{P' \in \mathbf{P}_\mathcal{C}} W_{d_\tau,1}(P, P')$ for any bucket order $\mathcal{C}$. In the case of Kendall's τ distance, it can be proven that $\Lambda_P(\mathcal{C}) = \sum_{i \prec_\mathcal{C} j} p_{j,i}$.

Dimensionality Reduction. Let $K \leq n$. We denote by $\mathbf{C}_K$ the set of all bucket orders $\mathcal{C}$ of $[\![n]\!]$ with K buckets. If P can be accurately approximated by a probability distribution associated to a bucket order with K buckets, a natural dimensionality reduction approach consists in finding a solution $\mathcal{C}^{*(K)}$ of

$$\min_{\mathcal{C} \in \mathbf{C}_K} \Lambda_P(\mathcal{C}), \tag{4}$$

as well as a solution $P^*_{\mathcal{C}^{*(K)}}$ of Eq. (3) for $\mathcal{C} = \mathcal{C}^{*(K)}$ and a coupling $(\Sigma, \Sigma_{\mathcal{C}^{*(K)}})$ s.t. $\mathbb{E}[d_\tau(\Sigma, \Sigma_{\mathcal{C}^{*(K)}})] = \Lambda_P(\mathcal{C}^{*(K)})$. Observe that $\cup_{\mathcal{C} \in \mathbf{C}_n} \mathbf{P}_\mathcal{C}$ is the set of all Dirac distributions δ_σ, $\sigma \in \mathfrak{S}_n$. Hence, in the case $K = n$, dimensionality reduction as formulated above boils down to solve *Kemeny consensus ranking* (see (2)), whereas the other extreme case $K = 1$ corresponds to no dimensionality reduction at all ($\Sigma_{\mathcal{C}^{*(1)}} = \Sigma$).

3 Our results

Firstly, we investigate the generalization capacity of minimizers of an empirical version of Eq. (4), i.e. based on realizations $\Sigma_1, \ldots, \Sigma_N$ drawn i.i.d. from P. Precisely, we obtained rate bounds for the excess risk of such solutions of order $\mathcal{O}_\mathbb{P}(1/\sqrt{N})$ for any distribution P and of order $\mathcal{O}_\mathbb{P}(1/N)$ under an additional assumption on the pairwise marginals of P. Furthermore, we demonstrate the relevance of our approach with experiments on real datasets, which show that one can keep a low distortion while drastically reducing the dimension of the distribution.

References

[1] J. G. Kemeny, 'Mathematics without numbers', *Daedalus*, **88**, 571–591, (1959).

[2] A. Korba, S. Clémençon, and E. Sibony, 'A learning theory of ranking aggregation', in *Proceeding of AISTATS 2017*, (2017).

[3] S.T. Rachev, *Probability Metrics and the Stability of Stochastic Models*, Wiley, 1991.

[1] Recall that a coupling of two probability distributions Q and Q' is a pair (U, U') of random variables defined on the same probability space such that the marginal distributions of U and U' are Q and Q'.