Clustering Rankings in the Fourier Domain

Stéphan Clémençon, Romaric Gaudel, Jérémie Jakubowicz

Clustering Rankings in the Fourier Domain

Stéphan Clémençon, Romaric Gaudel, and Jérémie Jakubowicz

LTCI, Telecom Paristech (TSI) - UMR Institut Telecom/CNRS No. 5141
stephan.clemencon@telecom-paristech.fr
romaric.gaudel@telecom-paristech.fr
jeremie.jakubowicz@telecom-paristech.fr

Abstract. It is the purpose of this paper to introduce a novel approach to clustering rank data on a set of possibly large cardinality $n \in \mathbb{N}^*$, relying upon Fourier representation of functions defined on the symmetric group $\mathfrak{S}_n$. In the present setup, covering a wide variety of practical situations, rank data are viewed as distributions on $\mathfrak{S}_n$. Cluster analysis aims at segmenting data into homogeneous subgroups, hopefully very dissimilar in a certain sense. Whereas considering dissimilarity measures/distances between distributions on the non commutative group $\mathfrak{S}_n$, in a coordinate manner by viewing it as embedded in the set $[0,1]^{n!}$ for instance, hardly yields interpretable results and leads to face obvious computational issues, evaluating the closeness of groups of permutations in the Fourier domain may be much easier in contrast. Indeed, in a wide variety of situations, a few well-chosen Fourier (matrix) coefficients may permit to approximate efficiently two distributions on $\mathfrak{S}_n$ as well as their degree of dissimilarity, while describing global properties in an interpretable fashion. Following in the footsteps of recent advances in automatic feature selection in the context of unsupervised learning, we propose to cast the task of clustering rankings in terms of optimization of a criterion that can be expressed in the Fourier domain in a simple manner. The effectiveness of the method proposed is illustrated by numerical experiments based on artificial and real data.

Keywords: clustering, rank data, non-commutative harmonic analysis, feature selection

1 Introduction

In a wide variety of applications, ranging from consumer relationship management (CRM) to economics through the design of recommendation engines for instance, data are often available in the form of rankings, *i.e.* partially ordered lists of objects expressing preferences (purchasing, investment, movies, *etc.*). Due to their global nature (modifying the rank of an object may affect that of many other objects), rank data generally deserve a special treatment in regards to statistical analysis. The latter have been indeed the subject of a good deal of attention in the machine-learning literature these last few years. Whereas numerous supervised learning algorithms have been recently proposed in order to

predict rankings accurately, see [FISS03], [PTA+07], [CV09] for instance, novel techniques have also been developed to handle input data, that are themselves of the form of rankings, for a broad range of purposes: computation of centrality measures such as consensus or median rankings (see [MPPB07] or [CJ10] for instance), modelling/estimation/simulation of distribution on sets of rankings (see [Mal57], [FV86], [LL03], [MM09] and [LM08] among others), ranking based on preference data (refer to [HFCB08], [CS01] or [dEW06]).

It is the main goal of this paper to consider the issue of clustering rank data from a novel perspective, taking the specific nature of the observations into account. We point out that the method promoted in this paper is by no means the sole possible applicable technique. The most widely used approach to this problem consists in viewing rank data (and, more generally, ordinal data), when renormalized in an appropriate manner, as standard numerical data and apply state-of-the-art clustering techniques, see Chapter 14 in [HTF09]. Alternative procedures of probabilistic type, relying on *mixture modeling* of probability distributions on a set of (partial) rankings, can also be considered, following in the footsteps of [FV88]. Rank data are here considered as probability distributions on the symmetric group $\mathfrak{S}_n$, $n \geq 1$ denoting the number of objects to be (tentatively) ranked and our approach crucially relies on the Fourier transform on the set of mappings $f : \mathfrak{S}_n \to \mathbb{R}$. Continuing the seminal contribution of [Dia89], spectral analysis of rank data has been recently considered in the machine-learning literature for a variety of purposes with very promising results, see [KB10], [HGG09] or [HG09]. This paper pursues this line of research. It aims at showing that, in the manner of spectral analysis in signal processing, Fourier representation is good at describing properties of distributions on $\mathfrak{S}_n$ in a sparse manner, "sparse" meaning here that a small number (compared to $n!$) of Fourier coefficients carry most of the significant information, from the perspective of clustering especially. As shown in [Dia88], the main appeal of spectral analysis in this context lies in the fact that Fourier coefficients encode structural properties of ranking distributions in a very interpretable fashion. Here we propose to use these coefficients as features for defining clusters. More precisely, we shall embrace the approach developed in [WT10] (see also [FM04]), in order to find clusters on an adaptively-chosen subset of features in the Fourier domain.

The article is organized as follows. Section 2 describes the statistical framework, set out the main notations and recall the key notions of Fourier representation in the context of distributions on the symmetric group that will be used in the subsequent analysis. Preliminary arguments assessing the efficiency of the Fourier representation for discrimination purposes are next sketched in Section 3. In particular, examples illustrating the capacity of parsimonious truncated Fourier expansions to approximate efficiently a wide variety of distributions on $\mathfrak{S}_n$ are exhibited. In Section 4, the rank data clustering algorithm we propose, based on subsets of spectral features, is described at length. Numerical results based on artificial and real data are finally displayed in Section 5. Technical details are deferred to the Appendix.

2 Background and Preliminaries

In this section, we recall key notions on spectral analysis of functions defined on the symmetric group $\mathfrak{S}_n$ and sketch an approximation framework based on this Fourier type representation, that will underly our proposed clustering algorithm.

2.1 Setup and First Notations

Here and throughout, $n \geq 1$ denotes the number of objects to be ranked, indexed by $i = 1, \ldots, n$. For simplicity's sake, it is assumed that no tie can occur in the present analysis, rankings being thus viewed as permutations of the list of objects $\{1, \ldots, n\}$ and coincide with the elements of the symmetric group $\mathfrak{S}_n$ of order n. Extension of the concepts developed here to more general situations (including *partial rankings* and/or *bucket orders*) will be tackled in a forthcoming article. The set of mappings $f : \mathfrak{S}_n \to \mathbb{C}$ is denoted by $\mathbb{C}[\mathfrak{S}_n]$. For any $\sigma \in \mathfrak{S}_n$, the function on $\mathfrak{S}_n$ that assigns 1 to σ and 0 to all $\tau \neq \sigma$ is denoted by δ_σ. The linear space $\mathbb{C}[\mathfrak{S}_n]$ is equipped with the usual inner product: $\langle f, g \rangle = \sum_{\sigma \in \mathfrak{S}_n} \overline{f(\sigma)} g(\sigma)$, for any $(f, g) \in \mathbb{C}[\mathfrak{S}_n]^2$. The related hilbertian norm is denoted by $||.||$. Incidentally, notice that $\{\delta_\sigma : \sigma \in \mathfrak{S}_n\}$ corresponds to the canonical basis of this Hilbert space. The indicator function of any event $\mathcal{E}$ is denoted by $\mathbb{I}\{\mathcal{E}\}$, the trace of any square matrix A with complex entries by $\mathrm{tr}(A)$, its conjugate by A^* and the cardinality of any finite set $\mathcal{S}$ by $\#\mathcal{S}$. Finally, for any $m \geq 1$, the matrix space $\mathcal{M}_{m \times m}(\mathbb{C})$ is equipped with the scalar product $< A, B >_m = \mathrm{tr}(A^* B)$ and the Hilbert-Schmidt norm $||A||_{HS(m)} = < A, A >_m^{1/2}$.

Rank Data as Distributions on $\mathfrak{S}_n$. The framework we consider stipulates that the observations are of the form of probability distributions on $\mathfrak{S}_n$, *i.e.* elements f of $\mathbb{C}[\mathfrak{S}_n]$ taking their values in $[0, 1]$ such that $\sum_{\sigma \in \mathfrak{S}_n} f(\sigma) = 1$. This models a variety of situations encountered in practice: indeed, list of preferences are rarely exhaustively observed, it is uncommon that rank data of the form of full permutations are available ($\sigma^{-1}(i)$ indicating the label of the i-th most preferred object, $i = 1, \ldots, n$). As shown by the following examples, a natural way of accounting for the remaining uncertainty is to model the observations as probability distributions. Precisely, let $\mathcal{E}_1, \ldots, \mathcal{E}_K$ be a partition of $\mathfrak{S}_n$, with $1 \leq K \leq n!$. When one observes which of these events is realized, *i.e.* in which of these subsets the permutation lies, the ensemble of possible observations is in one-to-one correspondence with the set of distributions $\{f_k : 1 \leq k \leq K\}$, where f_k denotes the conditional distribution of S given $S \in \mathcal{E}_k$, S being uniformly distributed on $\mathfrak{S}_n$, *i.e.* $f_k = (1/\#\mathcal{E}_k) \cdot \sum_{\sigma \in \mathcal{E}_k} \delta_\sigma$. In all these situations, the number of events K, through which preferences can be observed, may be very large.

Example 1. (Top-k lists.) In certain situations, only a possibly random number m of objects are ranked, those corresponding to the most preferred objects. In this case, the observations are related to the events of the type $\mathcal{E} = \{\sigma \in \mathfrak{S}_n :$

$(\sigma(i_1),\ \ldots, \sigma(i_m)) = (1,\ \ldots,\ m)\}$ where $m \in \{1,\ \ldots,\ n\}$ and $(i_1,\ \ldots,\ i_m)$ is a m-tuple of the set of objects $\{1,\ \ldots,\ n\}$.

Example 2. (PREFERENCE DATA.) One may also consider the case where a collection of objects, drawn at random, are ranked by degree of preference. The events observed are then of the form $\mathcal{E} = \{\sigma \in \mathfrak{S}_n :\ \sigma(i_1) < \ldots < \sigma(i_m)\}$ with $m \in \{1,\ \ldots,\ n\}$ and $(i_1,\ \ldots,\ i_m)$ a m-tuple of the set of objects $\{1,\ \ldots,\ n\}$.

Example 3. (BUCKET ORDERS.) The top-k list model can be extended the following way, in order to account for situations where preferences are *aggregated*. One observes a random partition $\mathcal{B}_1,\ \ldots,\ \mathcal{B}_J$ of the set of instances for which: for all $1 \leq j < l \leq J$ and for any $(i, i') \in \mathcal{B}_j \times \mathcal{B}_l$, $\sigma(i) < \sigma(i')$.

Hence, our objective is here to partition a set of N probability distributions $f_1,\ \ldots,\ f_N$ on $\mathfrak{S}_n$ into subgroups $\mathcal{C}_1,\ \ldots,\ \mathcal{C}_M$, so that the distributions in any given subgroup are closer to each other (in a sense that will be specified) than to those of other subgroups. When equipped with a *dissimilarity measure* $D : \mathfrak{S}_n \times \mathfrak{S}_n \to \mathbb{R}_+$, the clustering task is then classically cast in terms of minimization of the empirical criterion

$$\widehat{\mathcal{W}}(\mathcal{C}) = \sum_{m=1}^{M} \sum_{1 \leq i < j \leq n} D(f_i, f_j) \cdot \mathbb{I}\{(f_i, f_j) \in \mathcal{C}_m^2\}, \tag{1}$$

over the set of all possible partitions $\mathcal{C} = \{\mathcal{C}_m :\ 1 \leq m \leq M\}$ with $M \geq 1$ cells, the choice M requiring next the use of *model selection* techniques, see [TWH01] and the references therein. Beyond the fact that this corresponds to a NP-hard optimization problem for which acceptably good solutions can be computed after a reasonable lapse of time through a variety of (meta-)heuristics, the choice of a distance or a dissimilarity measure between pairs of distributions on $\mathfrak{S}_n$ is crucial. We emphasize that depending on the measure D chosen, one may either enhance specific patterns in the rank data or else make them disappear. This strongly advocates for the application of recent adaptive procedures that permit to achieve great flexibility in measuring dissimilarities, see [WT10] (refer also to [FM04]), selecting automatically a subset of attributes in order to emphasize differences between the cluster candidates. In this article, an attempt is made to implement this approach when attributes/features are those that are output by spectral analysis on $\mathbb{C}[\mathfrak{S}_n]$.

2.2 The Fourier Transform on $\mathfrak{S}_n$

Fourier analysis is an extraordinary powerful tool, whose usefulness and ubiquity is unquestionable and well documented, see [K$\ddot{\text{8}}$9,KLR95]: in signal and image processing it is used for the purpose of building filters and extracting information, in probability theory it permits to characterize probability distributions and serves as a key tool for proving limit theorems, in time-series analysis it allows to analyze second order stationary sequences, in analysis for solving partial

differential equations, *etc.* In fact, it is hard to think about a tool in (applied) mathematics that would be more widespread than the Fourier transform. The most common framework deals with functions $f : G \to \mathbb{C}$ where G denotes the group $\mathbb{R}$ of real numbers, the group $\mathbb{Z}$ of integers or the group $\mathbb{Z}/N\mathbb{Z}$ of integers modulo N. The elements of the abelian group G act on functions $f : G \to \mathbb{C}$ by *translation*. Recall that, for any $g \in G$, the translation by g is defined by $\mathcal{T}_g(f) : x \in G \mapsto f(x-g)$. A crucial property of the Fourier transform $\mathcal{F}$ is that it diagonalizes all translation operators simultaneously: $\forall g \in G$,

$$\mathcal{F}(\mathcal{T}_g(f))(\xi) = \chi_g(\xi) \cdot \mathcal{F}f(\xi),$$

where $\chi_g(\xi) = \exp(2i\pi g\xi)$ and ξ belongs to the dual group $\widehat{G}$ (being $\mathbb{R}$, $\mathbb{R}/\mathbb{Z}$ and $\mathbb{Z}/N\mathbb{Z}$ when G is $\mathbb{R}$, $\mathbb{Z}$ and $\mathbb{Z}/N\mathbb{Z}$ respectively). Consequently, Fourier transform provides a sparse representation of all operators that are spanned by the collection of translations, such as convolution operators.

The diagonalization view on Fourier analysis extends to the case of non-commutative groups such as $\mathfrak{S}_n$. However, in this case, the related eigenspaces are not necessarily of dimension 1 anymore. In brief, in this context Fourier transform only "block-diagonalizes" translations, as shall be seen below.

The Group Algebra $\mathbb{C}[\mathfrak{S}_n]$. The set $\mathbb{C}[\mathfrak{S}_n]$ is a linear space on which $\mathfrak{S}_n$ acts linearly as a group of translations $\mathcal{T}_\sigma : f \in \mathbb{C}[\mathfrak{S}_n] \mapsto \mathcal{T}_\sigma(f) = \sum_{\nu\in\mathfrak{S}_n} f(\nu \circ \sigma^{-1})\delta_\nu$. For clarity's sake, we recall the following notion.

Definition 1. (Convolution Product) *Let $(f, g) \in \mathbb{C}[\mathfrak{S}_n]^2$. The convolution product of g with f (with respect to the counting measure on $\mathfrak{S}_n$) is the function defined by $f * g : \sigma \in \mathfrak{S}_n \mapsto \sum_{\nu\in\mathfrak{S}_n} f(\nu)g(\nu^{-1} \circ \sigma) \in \mathbb{C}$.*

Remark 1. Notice that one may also write, for any $(f, g) \in \mathbb{C}[\mathfrak{S}_n]^2$ and all $\sigma \in \mathfrak{S}_n$, $(f * g)(\sigma) \sum_{\nu\in\mathfrak{S}_n} f(\sigma \circ \nu^{-1})g(\nu)$. The convolution product $f * \delta_\sigma = \tau \in \mathfrak{S}_n \mapsto f(\tau \circ \sigma^{-1})$ reduces to the right translation of f by σ, $\mathcal{T}_\sigma f$ namely. Observe in addition that, for $n > 2$, the convolution product is not commutative. For instance: $\delta_\sigma * \delta_\tau = \delta_{\sigma\circ\tau} \neq \delta_{\tau\circ\sigma} = \delta_\tau * \delta_\sigma$, when $\sigma \circ \tau \neq \tau \circ \sigma$.

The set $\mathbb{C}[\mathfrak{S}_n]$ equipped with the pointwise addition and the convolution product (see Definition 1 above) is referred to as the group algebra of $\mathfrak{S}_n$.

Canonical Decomposition. In the group algebra formalism introduced above, a function f is an eigenvector for *all* the right translations (simultaneously) whenever $\forall\sigma \in \mathfrak{S}_n$, $\delta_\sigma * f = \chi_\sigma f$, where $\chi_\sigma \in \mathbb{C}$ for all $\sigma \in \mathfrak{S}_n$. For instance, the function $f = \sum_{\sigma\in\mathfrak{S}_n} \delta_\sigma \equiv 1$ can be easily seen to be such an eigenvector with $\chi_\sigma \equiv 1$. In addition, denoting by ϵ_σ the signature of any permutation $\sigma \in \mathfrak{S}_n$ (recall that it is equal to $(-1)^{I(\sigma)}$ where $I(\sigma)$ is the number of inversions of σ, *i.e.* the number of pairs (i, j) in $\{1, \ldots, n\}^2$ such that $i < j$ and $\sigma(i) > \sigma(j)$), the function $f = \sum_{\sigma\in\mathfrak{S}_n} \epsilon_\sigma\delta_\sigma$ is also an eigenvector for all the right translations with $\chi_\sigma = \epsilon_\sigma$. If one could possibly find $n!$ such linearly independent eigenvectors, one would be able to define a notion of Fourier transform with properties very similar to those of the Fourier transform of functions defined on $\mathbb{Z}/N\mathbb{Z}$.

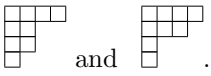
Unfortunately, due to the lack of commutativity of $\mathfrak{S}_n$, the functions mentioned above are the only eigenvectors common to all right translation operators, up to a multiplicative constant. Switching from the notion of eigenvectors to that of *irreducible subspaces* permits to define the Fourier transform, see [Ser88].

Definition 2. (Irreducible Subspaces) *A non trivial vector subspace V of $\mathbb{C}[\mathfrak{S}_\mathfrak{n}]$ (i.e. different from $\{0\}$ and $\mathbb{C}[\mathfrak{S}_\mathfrak{n}]$) is said to be irreducible when it is stable under all right translations, i.e. for all $(\sigma, f) \in \mathfrak{S}_n \times V$, $\delta_\sigma * f \in V$ and contains no such stable subspace except $\{0\}$ and itself. Two irreducible subspaces V_1 and V_2 are said to be isomorphic when there exists a bijective linear map $T : V_1 \to V_2$ that commutes with translations: $\forall f \in V_1$, $\forall \sigma \in \mathfrak{S}_n$, $T(\delta_\sigma * f) = \delta_\sigma * T(f)$.*

It follows from a well known result in Group Theory (in the compact case, see Peter-Weyl theorem in [Ser88] for instance) that $\mathbb{C}[\mathfrak{S}_\mathfrak{n}]$ decomposes into a direct sum of orthogonal irreducible subspaces. Spectral/harmonic analysis of an element $f \in \mathbb{C}[\mathfrak{S}_\mathfrak{n}]$ consists then in projecting the latter onto these subspaces and determine in particular which components contribute most to its "energy" $||f||^2$. In the case of the symmetric group $\mathfrak{S}_\mathfrak{n}$, the elements of the irreducible representation cannot be indexed by "scalar frequencies", the Fourier components being actually indexed by the set $\mathcal{R}_n$ of all *integer partitions of* n, namely:

$$\mathcal{R}_n = \left\{ \left\{ \xi = (n_1, \ \ldots, \ n_k) \in \mathbb{N}^{*k} : \ n_1 \geq \cdots \geq n_k, \sum_{i=1}^{k} n_i = n \right\}, 1 \leq k \leq n \right\}.$$

Remark 2. (Young tableaux) We point out that each element $(n_1, \ \ldots, \ n_k)$ of the set $\mathcal{R}_n$ can be visually represented as a Young tableau, with k rows and n_i cells at row $i \in \{1, \ \ldots, \ k\}$. For instance, the partition of $n = 9$ given by $\xi = (4, 2, 2, 1)$ and its conjugate ξ' can be respectively encoded by the diagrams:

 and .

Hence, the Fourier transform of any function $f \in \mathbb{C}[\mathfrak{S}_\mathfrak{n}]$ is of the form: $\forall \xi \in \mathcal{R}_n$,

$$\mathcal{F}f(\xi) = \sum_{\sigma \in \mathfrak{S}_n} f(\sigma)\rho_\xi(\sigma),$$

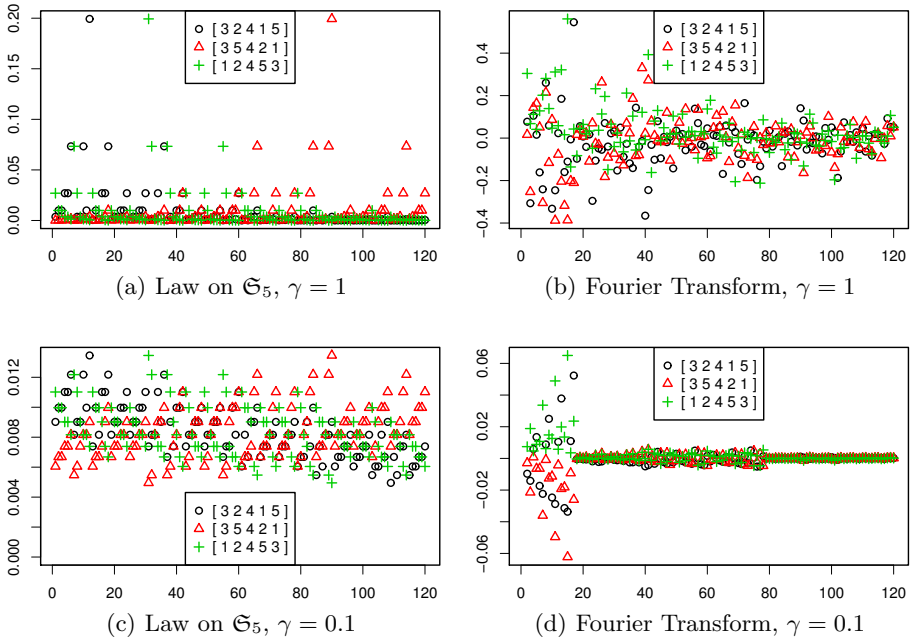
where ρ_ξ is a function on $\mathfrak{S}_\mathfrak{n}$ that takes its values in the set of unitary matrices with complex entries of dimension $d_\xi \times d_\xi$. Note that $\sum_\xi d_\xi^2 = n!$. For clarity, we recall the following result, that summarizes some crucial properties of the spectral representation on $\mathfrak{S}_n$, analogous to those of the standard Fourier transform. We refer to [Dia88] for an nice account of linear representation theory of the symmetric group $\mathfrak{S}_n$ as well as some of its statistical applications.

Proposition 1. (Main properties of $\mathcal{F}$) *The following properties hold true.*

(*i*) (Plancherel formula) $\forall (f, g) \in \mathbb{C}[\mathfrak{S}_n]^2$, $\langle f, g \rangle =< \mathcal{F}f, \mathcal{F}g >$, *where* $< \mathcal{F}f, \mathcal{F}g >= \frac{1}{n!} \sum_{\xi \in \mathcal{R}_n} d_\xi < \mathcal{F}f(\xi), \mathcal{F}g(\xi) >$.

(ii) (INVERSION FORMULA) $\forall f \in \mathbb{C}[\mathfrak{S}_n]$, $f = \frac{1}{n!}\sum_{\xi\in\mathcal{R}_n} d_\xi < \rho_\xi(.), \mathcal{F}f(\xi) >_{d_\xi}$.
(iii) (PARSEVAL FORMULA) $\forall f \in \mathbb{C}[\mathfrak{S}_n]$, $||f||^2 = \frac{1}{n!}\sum_{\xi\in\mathcal{R}_n} d_\xi ||\mathcal{F}f(\xi)||^2_{HS(d_\xi)}$.

Example 4. For illustration purpose, Fig. 1 below displays the spectral analysis of the Mallows distribution (*cf* [Mal57]) when $n = 5$, given by $f_{\sigma_0,\gamma}(\sigma) = \{\prod_{j=1}^{n}(1-\exp\{-\gamma\})/(1-\exp\{-j\gamma\})\}\cdot\exp\{-\gamma\cdot d_\tau(\sigma,\sigma_0)\}$ for all $\sigma \in \mathfrak{S}_n$, denoting by $d_\tau(\sigma,\nu) = \sum_{1\le i<j\le n}\mathbb{I}\{\sigma\circ\nu^{-1}(i) > \sigma\circ\nu^{-1}(j)\}$ the Kendall τ distance, for several choices of the location and concentration parameters $\sigma_0 \in \mathfrak{S}_n$ and $\gamma \in \mathbb{R}_+^*$. Precisely, the cases $\gamma = 0.1$ and $\gamma = 1$ have been considered. As shown by the plots of the coefficients, the more spread the distribution (*i.e.* the smaller γ), the more concentrated the Fourier coefficients.



(a) Law on $\mathfrak{S}_5$, $\gamma = 1$ (b) Fourier Transform, $\gamma = 1$

(c) Law on $\mathfrak{S}_5$, $\gamma = 0.1$ (d) Fourier Transform, $\gamma = 0.1$

Fig. 1. Coefficients of the Mallows distribution and of its Fourier transform for several choices of the pair of parameters (σ, γ).

Remark 3. (COMPUTATIONAL ISSUES) We point out that rank data generally exhibit a built-in high dimensionality. In many applications, the number of objects n to be possibly ranked is indeed very large. Widespread use of methods such as that presented here is conditioned upon advances in the development of practical fast algorithms for computing Fourier transforms. All the displays presented in this paper have been computed thanks to the C++ library $\mathbb{S}_n$ob, see [Kon06], that implements Clausen's fast Fourier transform for $\mathfrak{S}_n$.

3 Sparse Fourier Representations of Rank Data

In this section, an attempt is made to exhibit preliminary theoretical and empirical results, which show that spectral analysis of rank data can provide sparse and/or denoised representations of rank data in certain situations, that are useful for discrimination purposes.

Sparse Linear Reconstruction. Let F be a random element of $\mathbb{C}[\mathfrak{S}_\mathfrak{n}]$ and consider the following approximation scheme that consists in truncating the Fourier representation by retaining the coefficients with highest norm only in the reconstruction. It can be implemented in three steps, as follows. Let $K \in \{1, \ldots, n!\}$ be fixed.

1. Perform the Fourier transform, yielding matrix coefficients $\{\mathcal{F}F(\xi)\}_{\xi \in \mathcal{R}_n}$.
2. Sort the frequencies by decreasing order of magnitude of expected weighted norm of the corresponding matrix coefficients: $\xi_{(1)}, \ldots, \xi_{(K)}$, where

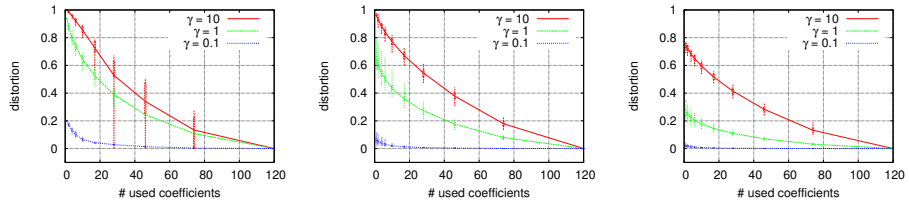
$$\mathbb{E}\left[d_{\xi_{(1)}} ||\mathcal{F}F(\xi_{(1)})||^2_{HS(d_{\xi_{(1)}})}\right] \geq \ \ldots \geq \mathbb{E}\left[d_{\xi_{(M)}} ||\mathcal{F}F(\xi_{(M)})||^2_{HS(d_{\xi_{(M)}})}\right],$$

   with $M = \#\mathcal{R}_n$, $d_{\xi_{(m)}}$ denoting the dimension of the irreducible space related to the frequency $\xi_{(m)}$.
3. Keeping the K coefficients with largest second order moment, invert the Fourier transform, producing the approximant

$$F_K(\sigma) = \frac{1}{n!} \sum_{k=1}^{K} d_{\xi_{(k)}} < \rho_{\xi_{(k)}}(\sigma), \mathcal{F}F(\xi_{(k)}) >_{d_{\xi_{(k)}}} . \tag{2}$$

The capacity of Fourier analysis to provide a sharp reconstruction of F can be evaluated through the expected distortion rate, namely $\epsilon_K(F) = \mathbb{E}[||F - F_K||^2]/\mathbb{E}[||F||^2]$. Fig. 2 shows the decay of a Monte-Carlo estimate (based on 50 runs) of the distortion rate $\epsilon_K(F)$ as the number K of retained coefficient grows, when F is drawn at random in a set of L Mallows distributions. As expected, the more spread the realizations of F (*i.e.* the smaller γ and/or the larger L), the faster the decrease (*i.e.* the more efficient the Fourier-based compression).

Remark 4. (On stationarity.) For simplicity, suppose that $\mathbb{E}[F(\sigma)]$ is constant, equal to 0 say. The expected distortion rate can be then easily expressed in terms of the covariance $\Phi(\sigma, \sigma') = \mathbb{E}[F(\sigma)F(\sigma')]$ and in the case when F is "stationary", *i.e.* when $\Phi(\sigma, \sigma') = \phi(\sigma\sigma'^{-1})$, it may be shown that Fourier representation has optimal properties in regards to linear approximation/compression, since the covariance operator is then diagonalized by the Fourier basis, exactly as in the well-known L_2 situation. Notice incidentally that such kernels $\phi(\sigma\sigma'^{-1})$ have already been considered in [Kon06] for a different purpose. Developing a full approximation theory based on Fourier representation of $\mathbb{C}[\mathfrak{S}_\mathfrak{n}]$, and defining in particular specific subspaces of $\mathbb{C}[\mathfrak{S}_\mathfrak{n}]$ that would be analogous to Sobolev classes and sparsely represented in the Fourier domain is beyond the scope of this paper and will be tackled in a forthcoming article.



(a) $L = 1$ (b) $L = 10$ (c) $L = 100$

Fig. 2. Monte-Carlo estimate of $\epsilon_K(F)$ as a function of K for a r.v. drawn at random in a set of L Mallows distributions with concentration parameter γ.

An Uncertainty Principle. The following proposition states an uncertainty principle in the context of the symmetric group. It extends the result established in [DS89] in the abelian setup to a non-commutative situation and can be viewed as a specific case of the inequality proved in [MS73] through operator-theoretic arguments. A simpler proof, directly inspired from the technique used in [DS89] in the commutative case, is given in the Appendix.

Proposition 2. (Uncertainty principle) *Let $f \in \mathbb{C}[\mathbb{S}_n]$. Denote by* $\mathrm{supp}(f) = \{\sigma \in \mathbb{S}_n : \ f(\sigma) \neq 0\}$ *and by* $\mathrm{supp}(\mathcal{F}f) = \{\xi \in \mathcal{R}_n : \ \mathcal{F}f(\xi) \neq 0\}$ *the support of f and that of its Fourier transform respectively. Then, we have:*

$$\#\mathrm{supp}(f) \cdot \sum_{\xi \in \mathrm{supp}(\hat{f})} d_\xi^2 \geq n!. \tag{3}$$

Roughly speaking, the inequality (3) above says in particular that, if the Fourier representation is sparse, *i.e.* $\mathcal{F}f(\xi)$ is zero at many frequencies ξ, $f(\sigma)$ is not.

De-noising in $\mathbb{C}[\mathfrak{S}_n]$. The following example shows how to achieve noise suppression through Fourier representation in some specific cases. Let $\mathcal{A} \neq \emptyset$ be a subset of $\mathfrak{S}_n$ and consider the uniform distribution on it: $f_\mathcal{A} = (1/\#\mathcal{A}) \cdot \sum_{\sigma \in \mathcal{A}} \delta_\sigma$. We suppose that it is observed with noise and shall study the noise effect in the Fourier domain. The noise is modeled as follows. Let T denote a transposition drawn at random in the set $\mathbf{T}$ of all transpositions in $\mathfrak{S}_n$ ($\mathbb{P}\{T = \tau\} = 2/(n(n-1))$ for all $\tau \in \mathbf{T}$), the noisy observation is: $f = f_A * \delta_T$. Notice that the operator modeling the noise here is considered in [RKJ07] for a different purpose.

We first recall the following result, proved by means of the Murnaghan-Nakayama rule ([Mur38]), that provides a closed analytic form for the expected Fourier transform of the random distribution f.

Proposition 3. *For all $\xi \in \mathcal{R}_n$, we have:*

$$\mathbb{E}[\mathcal{F}f(\xi)] = a_\xi \cdot \mathcal{F}f_\mathcal{A}(\xi), \tag{4}$$

where $a_\xi = (\sum_{i=1}^{r(\xi)} n_i(\xi)^2 - \sum_{i=1}^{r(\xi')} n_i(\xi')^2)/(n(n-1))$*, denoting by* $r(\xi)$ *the number of rows in the Young diagram representation of any integer partition of* n ξ *and* $n_i(\xi)$ *the number of cells at row* i*.*

The proposition above deserves some comments. Notice first that the map $\xi \in \mathcal{R}_n \mapsto a_\xi$ is antisymmetric (*i.e.* $a_{\xi'} = -a_\xi$ for any ξ in $\mathcal{R}_n$) and one can show that a_ξ is a decreasing function for the natural partial order on Young diagram [Dia88]. As it is shown in [RS08], $|a_\xi| = O(n^{-1/2})$ for diagrams ξ satisfying $r(\xi), c(\xi) = O(\sqrt{n})$. For the two lowest frequencies, (n) and $(n, n-1)$ namely, the proposition shows that $a_{(n)} = 1$ and $a_{(n-1,1)} = (n-3)/(n-1)$. Roughly speaking, this means that the noise leaves almost untouched (up to a change of sign) the highest and lowest frequencies, while attenuating moderate frequencies. Looking at extreme frequencies hopefully allows to recover/identify A.

Remark 5. (A MORE GENERAL NOISE MODEL) The model above can be extended in several manners, by considering, for instance, noisy observations of the form $f = f_{\mathcal{A}} * \delta_{S_m}$, where S_m is picked at random among permutations that can be decomposed as a composition of $m \geq 1$ transpositions (and no less). One may then show that $\mathbb{E}[\mathcal{F}f(\xi)] = a_\xi^{(m)} \cdot \mathcal{F}f_{\mathcal{A}}(\xi)$, for all $\xi \in \mathcal{R}_n$, where the $a_\xi^{(m)}$'s satisfy the following property: for all frequencies $\xi \in \mathcal{R}_n$ whose row number and column number are both less than $c_1\sqrt{n}$ for some constant $c_1 < \infty$, $|a_\xi^{(m)}| \leq c_2 \cdot n^{-m/2}$, where c_2 denotes some finite constant. See [RS08] for further details.

4 Spectral Feature Selection and Sparse Clustering

If one chooses to measure dissimilarity in $\mathbb{C}[\mathfrak{S}_n]$ through the square hilbertian norm, the task of clustering a collection of N distributions $f_1, \ldots, f_N$ on the symmetric group $\mathfrak{S}_n$ into $L << n!$ groups can be then formulated as the problem of minimizing over all partitions $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_L\}$ of the dataset the criterion:

$$\begin{aligned}\widehat{\mathcal{M}}(\mathcal{C}) &= \sum_{l=1}^{L} \sum_{1 \leq i,\, j \leq N} ||f_i - f_j||^2 \cdot \mathbb{I}\{(f_i, f_j) \in \mathcal{C}_l^2\} \\ &= \frac{1}{n!} \sum_{\xi \in \mathcal{R}_n} d_\xi \sum_{l=1}^{L} \sum_{1 \leq i,\, j \leq N:\, (f_i, f_j) \in \mathcal{C}_l^2} ||\mathcal{F}f_i(\xi) - \mathcal{F}f_j(\xi)||^2_{HS(d_\xi)},\end{aligned}$$

switching to the Fourier domain by using Parseval relation (see Proposition 1). Given the high dimensionality of rank data (*cf* Remark 3), such a rigid fashion of measuring dissimilarity may prevent the optimization procedure from identifying the clusters, the main features that are possibly responsible for the differences being buried in the criterion. As shown in the previous section, in certain situations only a few well-chosen spectral features may permit to exhibit similarities or dissimilarities between distributions on $\mathfrak{S}_n$. Following in the footsteps of [WT10], we propose to achieve a *sparse clustering* of the rank data by

considering the optimization problem: $\lambda > 0$ being a tuning parameter, minimize

$$\widehat{\mathcal{M}}_\omega(\mathcal{C}) = \sum_{\xi\in\mathcal{R}_n} \omega_\xi \frac{d_\xi}{n!} \sum_{l=1}^{L} \sum_{1\leq i,\, j\leq N:\, (f_i,f_j)\in\mathcal{C}_l^2} ||\mathcal{F}f_i(\xi) - \mathcal{F}f_j(\xi)||^2_{HS(d_\xi)} \quad (5)$$

$$\text{subject to } \omega = (\omega_\xi)_{\xi\in\mathbb{R}_n} \in \mathbb{R}_+^{\#\mathcal{R}_n},\ ||\omega||_{l_2}^2 \leq 1 \text{ and } ||\omega||_{l_1} \leq \lambda, \quad (6)$$

where $||\omega||_{l_2}^2 \overset{def}{=} \sum_{\xi\in\mathbb{R}_n} \omega_\xi^2$ and $||\omega||_{l_1} \overset{def}{=} \sum_{\xi\in\mathbb{R}_n} |\omega_\xi|$. The coefficient ω_ξ must be viewed as the weight of the spectral feature indexed by ξ. Observe incidentally that $(1/\sqrt{\#\mathcal{R}_n})\cdot\widehat{\mathcal{M}}(\mathcal{C})$ corresponds to the situation where $\omega_\xi = 1/\sqrt{\#\mathcal{R}_n}$ for all $\xi \in \mathcal{R}_n$. Following the *lasso* paradigm, the l_1-penalty guarantees the sparsity of the solution ω for small values of the threshold λ, while the l_2 penalty prevents the weight vector to concentrate on a single frequency, the one that most exhibits clustering.

Remark 6. (Further feature selection) Recall that the Fourier coefficients are matrices of variable sizes. Hence, denoting (abusively) by $\{\mathcal{F}f(\xi)_m :\ 1 \leq m \leq d_\xi\}$ the coefficients of the projection of any $f \in \mathbb{C}[\mathfrak{S}_n]$ onto the irreducible subspace V_ξ chosen in a given orthonormal basis of V_ξ, one may consider a refined sparse clustering, namely by considering the objective

$$\sum_{\xi\in\mathcal{R}_n} \sum_{m=1}^{d_\xi} \omega_{\xi,m} \sum_{l=1}^{L} \sum_{1\leq i,\, j\leq N:\, (f_i,f_j)\in\mathcal{C}_l^2} \left(\mathcal{F}f_i(\xi)_m - \mathcal{F}f_j(\xi)_m\right)^2.$$

Though it may lead to better results in practice, interpretation of large values of the weights as significant contributions to the clustering by the corresponding features becomes harder since the choice of a basis of V_ξ is arbitrary.

As proposed in [WT10], the optimization problem (5) under the constraints (6) is solved iteratively, in two stages, as follows. Starting with weights $\omega = (1/\sqrt{\#\mathcal{R}_n},\ \ldots,\ 1/\sqrt{\#\mathcal{R}_n})$, until convergence, iterate the two steps:

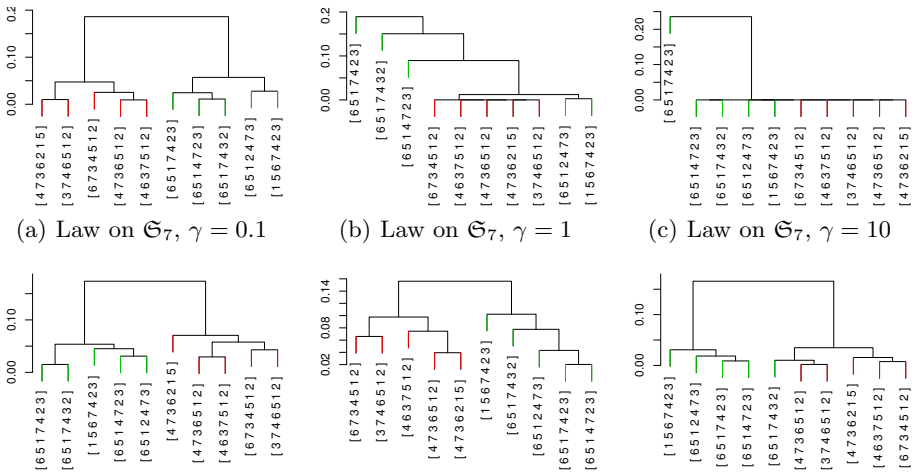
(Step 1.) fixing the weight vector ω, minimize $\widehat{\mathcal{M}}_\omega(\mathcal{C})$ after the partition $\mathcal{C}$,
(Step 2.) fixing the partition $\mathcal{C}$, minimize $\widehat{\mathcal{M}}_\omega(\mathcal{C})$ after ω.

A wide variety of clustering procedures have been proposed in the literature to perform Step 1 approximately (see [WX08] for an overview of off-the-shelf algorithms). As regards Step 2, it has been pointed out in [WT10] (see Proposition 1 therein), that a closed analytic form is available for the solution ω, the current data partition $\mathcal{C}$ being held fixed: $\omega_\xi = S(Z(\mathcal{C},\xi)_+, \Delta)/\|S(Z(\mathcal{C},\xi)_+, \Delta)\|_2$, for all $\xi \in \mathcal{R}_n$, where x_+ denotes the positive part of any real number x, $Z(\mathcal{C},\xi) = \sum_{l\neq l'} \sum_{(i,j):\, (f_i,f_j)\in\mathcal{C}_l\times\mathcal{C}_{l'}} ||\mathcal{F}f_i(\xi) - \mathcal{F}f_j(\xi)||^2_{HS(d_\xi)}$, and S is the *soft thresholding* function $S(x,\Delta) = sign(x)(|x| - \Delta)_+$, with $\Delta = 0$ if the l_1 constraint is fulfilled, and chosen positive so as to enforce $\|\boldsymbol{\omega}\|_{l_1} = s$ otherwise.

5 Numerical Experiments

This section presents experiments on artificial and real data which demonstrate that sparse clustering on Fourier representation recovers clustering information on rank data. For each of the three studied datasets, a hierarchical clustering is learned using [WT10] approach, as implemented in the `sparcl` R library. The clustering phase starts with each example in a separate cluster and then merges clusters two by two. Parameter s, controlling the l_1 norm of ω, is determined after the approach introduced in [WT10], which is related to the gap statistic used by [TWH01].

Hierarchical Clustering on Mallows Distributions. The first experiment considers an artificial dataset composed of $N = 10$ examples, where each example is a probability distribution on $\mathfrak{S}_n$. Formally, each example f_i corresponds to the Mallows distribution of center σ_i and spreading parameter γ. Centers σ_1 to σ_5 are of the form $\sigma^{(0)}\tau_i$ where $\sigma^{(0)}$ belongs to $\mathfrak{S}_n$ and τ_i is a transposition uniformly drawn in the set of all transpositions. Remaining centers follow a similar form $\sigma^{(1)}\tau_i$. Therefore examples should be clustered in two groups: one with the five first examples and a second group with remaining examples. Figure 3 gives the learned dendograms either from the distribution f_i, or from its Fourier transform $\mathcal{F}f_i$. The cluster information is fully recovered with the Fourier representation (except for one example), whereas the larger γ, the further from the target clustering is the dendogram learned from the distribution probabilities. Hence, the Fourier representation is more appropriate than the $[0, 1]^{n!}$ representation to cluster these rank data.



(a) Law on $\mathfrak{S}_7$, $\gamma = 0.1$ (b) Law on $\mathfrak{S}_7$, $\gamma = 1$ (c) Law on $\mathfrak{S}_7$, $\gamma = 10$

(d) Fourier, $\gamma = 0.1$ (e) Fourier, $\gamma = 1$ (f) Fourier, $\gamma = 10$

Fig. 3. Dendograms learned on $N = 10$ Mallows models on $\mathfrak{S}_7$ with spreading parameter γ. Examples which should be in the same cluster are plotted with same color.

Table 1. Number of selected coefficients, with $N = 10$ Mallows models on $\mathfrak{S}_7$ and spreading parameter γ.

Representation	# of coefficients			
	Total	Selected when $\gamma = 0.1$	Selected when $\gamma = 1$	Selected when $\gamma = 10$
Probability distribution	5,040	8	3	1
Fourier transform	5,040	257	54	6

Regarding the number of coefficients selected to construct these dendograms (*cf* Table 1), `sparcl` focuses on less coefficients when launched on the distribution probability. Still, with both representations, less than 260 coefficients are selected which remains small compare to the total number of coefficients (5,040). Unsurprisingly, in the Fourier domain, the selected coefficients depend on γ. When γ is small, the approach selects small frequency coefficients, whereas, when γ is large, the used coefficients are those corresponding to high frequencies.

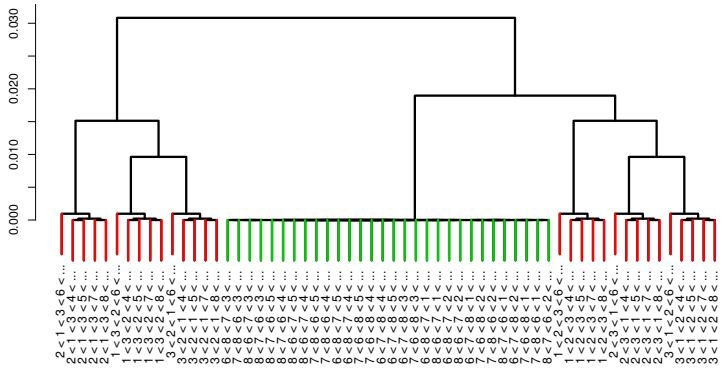


Fig. 4. Dendogram learned on top-k lists. Red lines (respectively green lines) correspond to examples $\mathcal{E}_i = \{\sigma \in \mathfrak{S}_8 : (\sigma_i(i_1), \ldots, \sigma_i(i_4)) = (1, \ldots, 4)\}$ with $\{i_1, i_2, i_3\} = \{1, 2, 3\}$ (resp. with $\{i_1, i_2, i_3\} = \{6, 7, 8\}$).

Hierarchical Clustering on Top-k Lists. This second experiment is conducted on artificial data, which consider rankings on eight products. Each example i of the data corresponds to a top-4 list $\mathcal{E}_i = \{\sigma \in \mathfrak{S}_8 : (\sigma_i(i_1), \ldots, \sigma_i(i_4)) = (1, \ldots, 4)\}$, where (i_1, i_2, i_3) is either a permutation on products $\{1, 2, 3\}$ or a permutation on products $\{6, 7, 8\}$, and i_4 is one of remaining products. All in all, this dataset contains 60 examples, which should be clustered in two groups or twelve groups. The 60 functions f_i have disjoint supports. Therefore, the examples cannot be clustered based on their "temporal" representation. On the contrary, the Fourier representation leads to a dendogram which groups together

examples for which the top-3 products are from the same three products (*cf* Figure 4). Furthermore, this dendogram is obtained from only seven small-frequency coefficients, which is extremely sparse compare to the total $40,320$ coefficients.

Hierarchical Clustering on an E-commerce Dataset. The proposed approach is also used to cluster the real dataset introduced by [RBEV10]. This data come from an E-commerce website. The only available information is the history of purchases for each user, and the ultimate goal is to predict future purchases. A first step to this goal is to group users with similar behavior, that means to group users based on the top-k rankings associated to their past purchases.

We consider the 149 users which have purchased at least 5 products among the 8 most purchased products. The sparse hierarchical clustering approach receives as input the 6,996 smallest frequency coefficients and selects 5 of them. The corresponding dendogram (*cf* Figure 5) clearly shows 4 clusters among the users. On 7 independent splits of the dataset in two parts of equal sizes, the criterion optimized by `sparcl` varies from 46.7 to 51.3 with a mean value of 49.1 and a standard deviation of 1.4. The stability of the criterion increases the confidence in this clustering of examples.

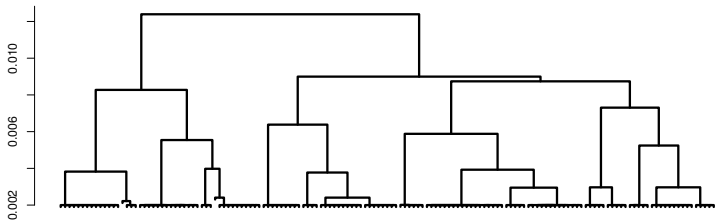


Fig. 5. Dendogram learned on the E-commerce database.

6 Conclusion

In this paper, a novel approach to rank data clustering is introduced. Modeling rank data as probability distributions on $\mathfrak{S}_n$, our approach relies on the ability of the Fourier transform on $\mathfrak{S}_n$ to provide sparse representations in a broad variety of situations. Several preliminary empirical and theoretical results are presented to support this claim. The approach to sparse clustering proposed in [WT10] is adapted to this setup: a lasso-type penalty is used so as to select adaptively a subset of spectral features in order to define the clustering. Numerical examples are provided, illustrating the advantages of this new technique. A better understanding of the class of distributions on $\mathfrak{S}_n$ that can be efficiently described in the Fourier domain and of the set of operators that are almost diagonal in the Fourier basis would permit to delineate the compass of this approach. It is our intention to develop this line of research in the future.

Appendix - Proof of Proposition 2

Combining the definition of $\mathcal{F}f$, triangular inequality and the fact that the ρ_ξ's take unitary values (in particular, $||\rho_\xi(.)||_{HS(d_\xi)} = \sqrt{d_\xi}$), we have: $\forall \xi \in \mathcal{R}_n$,

$$\begin{aligned} ||\mathcal{F}f(\xi)||_{HS(d_\xi)} &= \sum_{\sigma \in \mathfrak{S}_n} f(\sigma)\rho_\xi(\sigma) \leq \sqrt{d_\xi}||f||_1 \\ &\leq \sqrt{d_\xi}(\#\text{supp}(f))^{1/2}||f|| \\ &\leq \sqrt{d_\xi}(\#\text{supp}(f))^{1/2} \left(\frac{1}{n!} \sum_{\xi \in \text{supp}(\mathcal{F}f)} d_\xi ||\mathcal{F}f(\xi)||^2_{HS(d_\xi)} \right)^{1/2}, \end{aligned}$$

where the two last bounds result from Cauchy-Schwarz inequality and Plancherel relation respectively, with $||f||_1 \stackrel{def}{=} \sum_{\sigma \in \mathfrak{S}_n} |f(\sigma)|$. Now, the desired bound immediately follows.

Acknowledgements. Authors greatly thank DIGITÉO (BÉMOL project), which partially supported this work.

References

[CJ10] S. Clémençon and J. Jakubowicz. Kantorovich distances between rankings with applications to rank aggregation. In *Proceedings of ECML'10*, 2010.

[CS01] K. Crammer and Y. Singer. Pranking with ranking. In *NIPS*, 2001.

[CV09] S. Clémençon and N. Vayatis. Tree-based ranking methods. *IEEE Transactions on Information Theory*, 55(9):4316–4336, 2009.

[dEW06] M. desJardins, E. Eaton, and K. Wagstaff. Learning user preferences for sets of objects. In *Proceedings of ICML'06*, pages 273–280, 2006.

[Dia88] P. Diaconis. *Group representations in probability and statistics*. Institute of Mathematical Statistics, Hayward, 1988.

[Dia89] P. Diaconis. A generalization of spectral analysis with application to ranked data. *The Annals of Statistics*, 17(3):949–979, 1989.

[DS89] D. Donoho and P. Stark. Uncertainty principles and signal recovery. *SIAM J. Appl. Math.*, 49(3):906–931, 1989.

[FISS03] Y. Freund, R. D. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *JMLR*, 4:933–969, 2003.

[FM04] J. H. Friedman and J. J. Meulman. Clustering objects on subsets of attributes. *JRSS*, 66(4):815–849, 2004.

[FV86] M. A. Fligner and J. S. Verducci. Distance based ranking models. *JRSS Series B (Methodological)*, 48(3):359–369, 1986.

[FV88] M. A. Fligner and J. S. Verducci. Multistage ranking models. *JASA*, 83(403):892–901, 1988.

[HFCB08] E. Hüllermeier, J. Fürnkranz, W. Cheng, and K. Brinker. Label ranking by learning pairwise preferences. *Artificial Intelligence*, 172:1897–1917, 2008.

[HG09] J. Huang and C. Guestrin. Riffled independence for ranked data. In *Proceedings of NIPS'09*, 2009.

[HGG09] J. Huang, C. Guestrin, and L. Guibas. Fourier theoretic probabilistic inference over permutations. *JMLR*, 10:997–1070, 2009.

[HTF09] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning (2nd ed.)*, pages 520–528. Springer, 2009.

[K̈89] T. Körner. *Fourier Analysis*. Cambridge University Press, 1989.

[KB10] R. Kondor and M. Barbosa. Ranking with kernels in Fourier space. In *Proceedings of COLT'10*, 2010.

[KLR95] J.P. Kahane and P.G. Lemarié-Rieusset. *Fourier series and wavelets*. Routledge, 1995.

[Kon06] R. Kondor. $\mathbb{S}_n$*ob: a C++ library for fast Fourier transforms on the symmetric group*, 2006. Available at http://www.its.caltech.edu/ risi/Snob/.

[LL03] G. Lebanon and J. Lafferty. Conditional models on the ranking poset. In *Proceedings of NIPS'03*, 2003.

[LM08] G. Lebanon and Y. Mao. Non-parametric modeling of partially ranked data. *JMLR*, 9:2401–2429, 2008.

[Mal57] C. L. Mallows. Non-null ranking models. *Biometrika*, 44(1-2):114–130, 1957.

[MM09] B. Mandhani and M. Meila. Tractable search for learning exponential models of rankings. In *Proceedings of AISTATS'09*, 2009.

[MPPB07] M. Meila, K. Phadnis, A. Patterson, and J. Bilmes. Consensus ranking under the exponential model. In *Proceedings of UAI'07*, pages 729–734, 2007.

[MS73] T. Matolcsi and J. Szücs. Intersection des mesures spectrales conjuguées. *CR Acad. Sci. S r. I Math.*, (277):841–843, 1973.

[Mur38] F. D. Murnaghan. *The Theory of Group Representations*. The Johns Hopkins Press, 1938.

[PTA+07] T. Pahikkala, E. Tsivtsivadze, A. Airola, J. Boberg, and T. Salakoski. Learning to rank with pairwise regularized least-squares. In *Proceedings of SIGIR'07*, pages 27–33, 2007.

[RBEV10] E. Richard, N. Baskiotis, T. Evgeniou, and N. Vayatis. Link discovery using graph feature tracking. In *NIPS'10*, pages 1966–1974, 2010.

[RKJ07] A. Howard R. Kondor and T. Jebara. Multi-object tracking with representations of the symmetric group. In *Proceedings og ICML'07*, 2007.

[RS08] A. Rattan and P. Sniady. Upper bound on the characters of the symmetric groups for balanced Young diagrams and a generalized Frobenius formula. *Adv. in Math.*, 218(3):673–695, 2008.

[Ser88] J. P. Serre. *Algebraic groups and class fields*. Springer-Verlag, NY, 1988.

[TWH01] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *J. Royal Stat. Soc.*, 63(2):411–423, 2001.

[WT10] D. M. Witten and R. Tibshirani. A framework for feature selection in clustering. *JASA*, 105(490):713–726, 2010.

[WX08] D. Wünsch and R. Xu. *Clustering*. IEEE Press, Wiley, 2008.