# Multiresolution analysis of ranking data

Eric Sibony

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2016-ENST-0036

EDITE - ED 130

# Doctorat ParisTech

# T H È S E

**pour obtenir le grade de docteur délivré par**

# TELECOM ParisTech

## Spécialité « Signal et Image »

*présentée et soutenue publiquement par*


## Éric SIBONY


le 14 juin 2016

# Analyse multirésolution de données de classements

Directeur de thèse : **Stéphan CLÉMENÇON**
Co-encadrement de la thèse : **Jérémie JAKUBOWICZ**

**Jury**

| | |
|---|---|
| **M. Stéphane MALLAT**, Professeur, École Normale Supérieure | Président |
| **M. Risi KONDOR**, Assistant Professor, The University of Chicago | Rapporteur |
| **M. Devavrat SHAH**, Associate Professor, MIT | Rapporteur |
| **M. Jonathan HUANG**, Research Scientist, Google | Examinateur |
| **M. Eyke HÜLLERMEIER**, Professor, Universität Paderborn | Examinateur |
| **Mme Michèle SEBAG**, Directeur de Recherche CNRS, Université Paris Sud | Examinateur |
| **M. Stéphan CLÉMENÇON**, Professeur, Télécom Paristech | Directeur de thèse |
| **M. Jérémie JAKUBOWICZ**, Maître de Conférence, Télécom SudParis | Co-encadrant |


**TELECOM ParisTech**
école de l'Institut Mines-Télécom - membre de ParisTech
46 rue Barrault 75013 Paris - (+33) 1 45 81 77 77 - www.telecom-paristech.fr


T H È S E

# Acknowledgments

I have often been asked what doing a PhD in mathematics really consisted in. If it seems quite difficult to give an exhaustive answer, I think one major aspect is that it consists in a subtle blend between solitary (even sometimes introspective) work and discussions or reflections with other people. I have indeed spent a large part of my time fighting with calculations, to end up with the solution after the $10^{th}$ attempt, then realizing there was a mistake and redoing it all; trying to understand the implication of some equation and realizing after one month that it was obvious; or tackling a problem for days to finally wonder "what is it that I am really doing?". Though this iterative and highly nonlinear process was certainly inevitable or even necessary to end up with some contributions, it would have been quite inefficient without the many enriching and eye-opening discussions I was lucky to have. So please allow me now to express my gratitude.

Mes premiers remerciements vont à mes directeurs de thèse Stéphan et Jérémie. Ce fut un réel plaisir de faire ma thèse avec vous. Je vous remercie d'abord pour m'avoir laissé la liberté de partir sur des pistes floues et d'amener le sujet sur des domaines imprévus, tout en m'accompagnant et me poussant toujours plus loin dans ce fantastique voyage. Au cours de ces années vous m'avez enseigné à toujours chercher à distinguer ce qui repose sur des grands principes mathématiques de ce qui tient à la structure spécifique des objets en présence. Stéphan, en me formant à la rédaction d'articles scientifiques, en m'envoyant à de nombreuses conférences et en me faisant découvrir les rouages du monde de la recherche, tu m'as fait devenir, je crois, un jeune chercheur mature. Tu m'as également fait comprendre que ma culture cinématographique était encore largement restreinte et qu'il me restait beaucoup de chemin à parcourir. Jérémie, au cours de nos séances feutre en main, tu m'as appris à toujours pousser plus loin la compréhension des objets mathématiques et algorithmes considérés. Les combiner avec toutes nos autres discussions fut de plus un réel plaisir. Pour tout cela et pour tout ce que je n'ai pas mentionné dans ces trop courtes lignes, merci.

Then I would like to thank the members of my PhD committee. I have been largely inspired by your work and it is a great honor as well as a great pleasure that you all accepted to be part of it. First to Risi Kondor and Devavrat Shah I express my gratitude for making me the honor to review my thesis. The interest you take in my work is a great encouragement to proceed on the subject. To Stéphane Mallat, Michèle Sebag and Eyke Hüllermeier, I am very happy I had the opportunity to discuss with you on various scientific subjects, it was always a great source of inspiration for me. To Jonathan Huang, I had not have the opportunity to meet you in person before but having spent so much time on your papers and your thesis, it is a great pleasure to finally get your live feedback.

I would now like to thank all the people I had the chance to meet during my PhD. First to Persi Diaconis, it goes without saying that your work was a prominent source of inspiration for me. Having the opportunity to discuss with you about my work was thus a great honor and a great pleasure. I sincerely hope this opportunity will present again in the future. To Gérard

Kerkyacharian, Aurélien Garivier, Albert Cohen, Hugues Randriambololona and David Madore, I thank you for your feedback, it was very valuable and helped me in constructing my work.

As a last source of inspiration, I would like to address a peculiar acknowledgment: to the Google search engine. Indeed, when I started to consider the space $H_{[\![n]\!]}$ (see Definition 40), I had the intuition that its dimension should be equal to the number of fixed-point free permutations of a set of $n$ elements, but I had no idea of how to prove it for any $n$. After one month of struggle, I asked Google if it knew about "a vector space with dimension equal to the number of fixed-point free permutations of a set of $n$ elements". It gave me references to articles from algebraic topology that were completely abstruse to me at first sight, but which eventually lead me to the result I needed (Theorem 51).

Pour remonter à l'origine de cette thèse, je dois bien sûr remercier mes professeurs de lycée et de prépa qui m'ont formé aux mathématiques : Bernard Randé, Serge Dupont et Jean-Pierre Sanchez. J'adresse un remerciement particulier à vous M. Sanchez, vous qui m'avez le premier appris les mathématiques, la rigueur du raisonnement et l'amour des jolies équations. C'est grâce à votre enseignement que j'ai décidé de devenir un scientifique. Je vous en serai toujours reconnaissant.

Sur un plan plus personnel maintenant, j'aimerais d'abord adresser mes remerciements à toute l'équipe STA de Télécom, en particulier les doctorants et post-docs des différentes générations que j'ai pu cotôyées. Ce fut un réel plaisir de vivre au sein de cette équipe et de partager autant de moments sympathiques, à Télécom, aux conférences ou ailleurs. Pour rendre à César, je me dois de remercier Claire dont les nombreuses initiatives ont permis de relancer cette si belle dynamique d'équipe qui je l'espère continuera de perdurer longtemps. A Anna, merci pour tes nombreux retours sur ma thèse. C'est un plaisir de continuer de travailler sur le sujet avec toi, plein de belles choses sont à venir j'en suis sûr.

Ensuite, je veux bien sûr remercier mes amis et associés Jeremy Jawish et David Durrlemann, ainsi que toute l'équipe de Shift. Vous m'avez toujours soutenu pendant ces quatre années de double vie, et jamais fait ressentir la moindre culpabilité. Bien sûr de mon point de vue il y a plein de choses que je n'ai pas pu faire, j'aurai maintenant le temps de rattraper le retard !

A tous mes amis, je vous remercie pour tous ces moments passés ensemble et je m'excuse pour toutes les fois où je n'ai pas pu répondre présent ou rendre la pareille à vos invitations parce que j'avais trop de travail. Ce sera dorénavant différent !

A mes parents et à mes frères enfin, je vous remercie d'abord pour votre aide dans la relecture de mon manuscrit et dans la préparation du pot. Mais plus généralement bien sûr, pour votre soutien sans faille depuis toutes ces années, vos nombreux conseils, votre aide continuelle, pour tous nos moments de vie et de joie et pour tout le reste, merci. Je vous aime.

# Contents

# Chapter 1

# Introduction

*"With four parameters I can fit an elephant, and with five I can make him wiggle his trunk."*

John Von Neumann

By this joke, John Von Neumann meant that with enough parameters one can fit any dataset; for all that, the resulting model does not necessarily represent the underlying data generating process well. In modern terms, one would say that John Von Neumann's elephant is overfitted with five parameters. How many parameters should be used to describe the world then? The question remains at the heart of all the data analysis literature.

More precisely, it has become common sense that the number of parameters should be small, turning the question into: what *representation* of the data leads to models with the fewest parameters? This approach is well exemplified in sparsity methods, where one approximates any function with a linear combination of a small number of "atoms" from a "dictionary". The goal is then to come up with the dictionary that yields the best sparse approximations of a large class of functions. Focus is thus made on the representation that the dictionary provides for the data rather than on the models, which have a simple expression in the representation. Many contributions have therefore introduced representations for a wide variety of applications. With the impressive success of deep learning, numerous recent methods even go a step further and seek to learn directly the representation from the data.

In this thesis, we introduce a new representation for ranking data. Much less considered than vector data, they actually arise in a large variety of applications and a tremendous literature is dedicated to their analysis, spreading across many scientific fields such as social choice theory, psychometry, statistics, economics, operations research, artificial intelligence or machine learning. Most approaches involve however "parametric" modeling of probability distributions over rankings. Though they lead to satisfying results, they lack flexibility: by essence, parametric models make an assumption on the data and do not enable to interpret outside of it. Some contributions have therefore developed "nonparametric" frameworks to enable general interpretation of the data and design inference methods based on regularity assumptions. Examples include Fourier analysis on the symmetric group, introduced in Diaconis (1988), or the HodgeRank framework, introduced in Jiang et al. (2011b) (refer to Chapter 6 for connections with the present work). Both are however fitted for a specific type of ranking data, full and partial rankings for the former and pairwise comparisons for the latter (see Chapter 2 for the definitions).

The representation we introduce is fitted for incomplete rankings. Generalizing pairwise comparisons, such rankings have a canonic multiscale structure. Our representation is thus naturally analogous to a *multiresolution analysis*. First formalized in Mallat (1989), multiresolution analysis has led to a tremendous number of applications in statistics and signal processing, and has been extended in many ways with remarkable success. Recent literature has dedicated a special interest for multiresolution analyses on non-vector data (Coifman and Maggioni, 2006; Gavish et al., 2010; Hammond et al., 2011; Rustamov and Guibas, 2013; Kondor et al., 2014) and the first multiresolution analysis on the symmetric group was introduced in Kondor and Dempsey (2012). Our work is of course inspired by all these contributions. It relies however on a very different mathematical construction and involves a specific notion of *information localization* developed at length in the thesis.

Our representation provides several new insights on ranking data analysis and offers a flexible framework to design efficient statistical procedures. We precise however that we do not introduce a generic method that outperforms the state-of-the-art in all applications. In particular we present numerical experiments with the naive application of the representation, which outperforms the state-of-the-art in small-scale settings but is dominated by classic parametric approaches in large-scale ones (see Chapter 5). This is no surprise because the representation does not make any assumption on the data, leading to a naive application with an important number of parameters. It therefore has a much larger variance than estimators based on parametric models with few parameters. By contrast, our representation offers a great design flexibility and, combined with the adapted regularization procedure for each application, should provide better inference methods for large-scale settings. This is why we provide a global survey of the ranking data analysis literature (Chapter 2) and detail numerous future research directions (Chapter 7).

From a general point of view, I believe that ranking data analysis will continue to know many new developments and applications, and will benefit from the advances of numerous mathematical areas. I hope that the present work will help in that trend.

**Outline.** The thesis is organized as follows.

- We first provide a global survey of ranking data analysis in Chapter 2 and introduce the general notations for this thesis.
- Then in Chapter 3 we formalize the general setting for the analysis of incomplete and detail at length the motivations for the present work.
- Chapter 4 is certainly the heart of this thesis: it introduces our new representation and develops in details its properties, algorithms, and associated interpretations.
- In Chapter 5 we introduce a general framework to use the representation for the statistical analysis of incomplete rankings and present some applications.
- Several connections with other mathematical constructions are established in Chapter 6, in particular with Fourier analysis and Kemeny rank aggregation.
- At last Chapter 7 contains the informal description of several future directions, and a general conclusion to this thesis.

**Related articles.** This thesis has lead or will lead to several papers. First, the core content was developed in the unpublished manuscripts Clémençon et al. (2014) and Sibony et al. (2016). Second, several specific parts were contained in the following papers accepted in conferences.

- Section 5.3 is based on Sibony et al. (2014)

- Most results of Subsections 6.3.2 and 6.3.3 were present in Sibony (2014)
- Section 5.2 and some results of Section 4.2 are from Sibony et al. (2015)

At last, we plan to submit the following articles to different journals.

- A survey on ranking data analysis based on Chapter 2, planned to be submitted to *Foundations and Trends in Machine Learning*.
- An article that details the construction of the MRA representation, based on Chapter 4 and Section 3.2, planned to be submitted to the *SIAM Journal on Discrete Mathematics*.
- An article for the statistical analysis of incomplete rankings, based on Chapter 5 and Section 3.1, planned to be submitted to the *Journal of Machine Learning Research*.
- An article about the connections with Fourier analysis and other constructions, based on Chapter 6, planned to be submitted to *Applied and Computational Harmonic Analysis*.

# Part I

# Background and motivations

# Chapter 2

# Ranking data analysis

We begin this thesis with a general overview of ranking data analysis. First we introduce the definitions and notations that will be used throughout the thesis. Then in Section 2.2 we describe the major domains of application where ranking data is analyzed. In Section 2.3, we formalize the main problems of ranking data analysis. We then use Section 2.4 to detail the specificities of these problems and the associated challenges. At last we describe in Section 2.5 the models that have been introduced in the literature to tackle these issues. We precise that because of the great importance of the ranking data analysis literature, our description is certainly not exhaustive.

## Contents

## 2.1 Modeling, definitions and notations

We first introduce the main definitions and notations that we use in this thesis. We also try at the most to develop interpretation for the mathematical objects we consider.

### 2.1.1 General definition

Ranking data represent ordinal comparisons. They model for instance the preferences of customers on different products, the ordering of participants in the results of a competition or the relative performance of several methods in different experiments (examples of applications are detailed in Subsection 2.2). From a mathematical point of view, comparisons are made between elements of a set. The first major dichotomy in the literature concerns the model for the set: is it finite or infinite? In the latter case, elements are usually characterized by a list of features so that the set is equal to a product space, typically $\mathbb{R}^d$. In the former case, elements are usually given an identifier, a number between 1 and $n$, the total number of elements, which solely characterizes them. The choice between the two models usually corresponds to the considered application and leads to specific mathematical developments. Many models for ranking data analysis can however be applied to both settings.

In this thesis we mainly consider the case where the set of elements is finite and each element is solely characterized by its identifier. Applications to an infinite set where each element is characterized by features are discussed in Chapter 7. Here and throughout the thesis, the set is denoted by $[\![n]\!] := \{1, \dots, n\}$, where $n \geq 1$ is the total number of elements.

**Definition 1** (Ranking)**.** A ranking is a strict partial order on $[\![n]\!]$: a nonempty collection of pairwise comparisons $a \prec b$ with $a, b \in [\![n]\!]$ that satisfies the following properties (the last property is implied by the first two).

- **Irreflexivity:** for all $a \in [\![n]\!]$, $a \not\prec a$
- **Transitivity:** for all $a, b, c \in [\![n]\!]$, if $a \prec b$ and $b \prec c$ then $a \prec c$
- **Asymmetry:** for all $a, b \in [\![n]\!]$, if $a \prec b$ then $b \not\prec a$

We denote by $\mathfrak{R}_n$ the set of all rankings on $[\![n]\!]$.

The reader can refer for instance to Stanley (1986) for more details about strict partial orders. By convention, $a \succ b$ means that element $a$ is preferred to or ranked higher than element $b$. Though a strict partial order is defined mathematically as a collection of pairwise comparisons, we will usually characterize it by an expression involving the following short notations: for $a, b, c \in [\![n]\!]$, $a \succ b, c$ means that $a \succ b$ and $a \succ c$, and $a \succ b \succ c$ means by transitivity that $a \succ b$, $b \succ c$ and $a \succ c$. One can easily represent a ranking on $[\![n]\!]$ by its *Hasse diagram*: each element of $[\![n]\!]$ is a node, and an arrow is drawn from element $a$ to element $b$ if $a \succ b$ and there is no element $c$ such that $a \succ c \succ b$. By convention, elements ranked higher are placed above in the diagram. Table 2.1 shows the Hasse diagram of all the possible rankings on $[\![n]\!]$ for $n = 3$.

*Remark* 2 (Modeling of ties). Definition 1 of a ranking enables to model ties in the following way: if some elements $a_1, \dots, a_k \in [\![n]\!]$ are all compared with the same outcome to an element $b \in [\![n]\!]$ but no comparison is given between them, they have equal rank. For instance in the

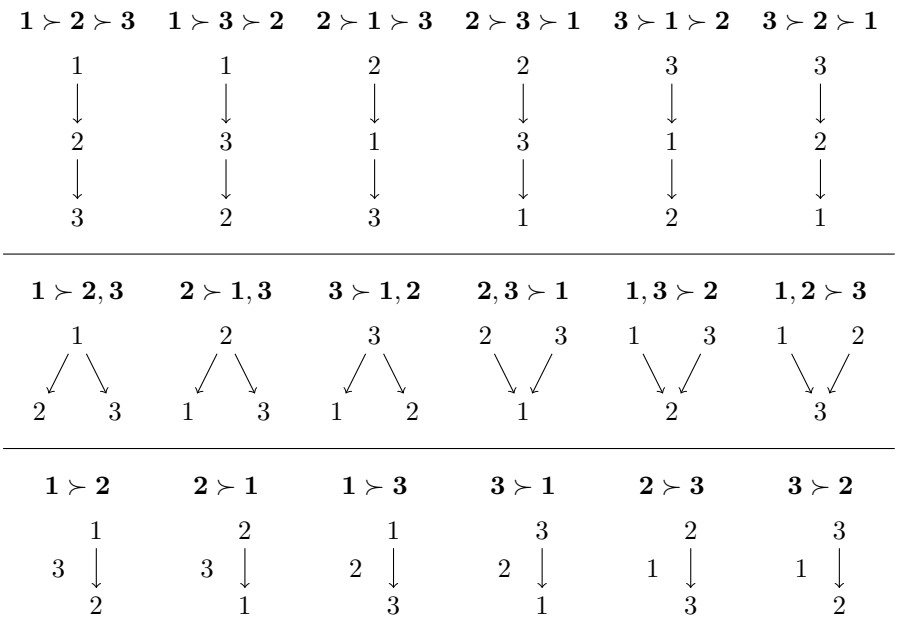



Table 2.1: Possible rankings on $[\![3]\!]$ and their Hasse diagrams

ranking $1 \succ 2, 3$ on $[\![3]\!]$, the element 1 is placed first and the elements 2 and 3 are placed equal second. We point out that this notion of tie is relative to the elements involved in the comparison. For instance in the ranking $1 \succ 2, 3$ and $4, 5 \succ 6$ on $[\![6]\!]$, the elements 2 and 3 are placed equal second with respect to the element 1 and the elements 4 and 5 are placed equal first with respect to the element 6. In particular, Definition 1 of a ranking does not enable to represent the sole observation of a tie between two elements. This possibility could be useful in some applications (to represent a tie in a sports game for instance) but taking it into account leads to specific mathematical developments. While some contributions tackle this issue (see for instance Rao and Kupper, 1967; Davidson, 1970; Batchelder and Bershad, 1979), most of the literature consider rankings in the scope of Definition 1.

### 2.1.2 Classes of rankings

The rankings on $[\![n]\!]$ are quite heterogeneous objects, and studying them in a general framework is very complex. Fortunately, some subclasses of rankings are much more homogeneous, in the sense that their elements share common properties. Contributions in the literature thus usually focus on one subclass and represents the rankings by equivalent but more practical mathematical objects. Some particular subclasses of rankings have attracted most of the attention. They each correspond to a row in Table 2.1 for $n = 3$, and we now detail their formal definitions. The vocabulary we use is, up to our knowledge, the most classic one in the literature (see Marden, 1996; Alvo and Yu, 2014).

**Full rankings.** Arguably the most studied subclass of rankings are *full rankings*, of the form

$$a_1 \succ \cdots \succ a_n,$$

where $a_1$ is the element of $[\![n]\!]$ ranked first and $a_n$ is the element of $[\![n]\!]$ ranked last. A full ranking corresponds to a *total order* $\prec$ on $[\![n]\!]$: for any distinct elements $a, b \in [\![n]\!]$, either $a \succ b$ or $a \prec b$. It is also called a *linear order*, because its Hasse diagram is a chain involving all the elements of $[\![n]\!]$, namely $a_1 \to \cdots \to a_n$. Full rankings on $[\![3]\!]$ are represented by the first row in Table 2.1. There is also a one-to-one correspondence between full rankings and permutations of $[\![n]\!]$ that is to say bijective mappings $\sigma : [\![n]\!] \to [\![n]\!]$. More specifically, we associate the full ranking $a_1 \succ \cdots \succ a_n$ with the permutation $\sigma$ that maps an element to its rank in the ranking: $\sigma$ is defined by $\sigma(a_i) = i$ for $i = 1, \ldots, n$. With this correspondence, $a \succ b$ is equivalent to $\sigma(a) < \sigma(b)$. The set of permutations of $[\![n]\!]$ is called the symmetric group and denoted by $\mathfrak{S}_n$. We do not distinguish between a full ranking and its associated permutation thereafter.

**Partial rankings.** A first generalization of full rankings are *partial rankings*, of the form

$$a_{1,1}, \ldots, a_{1,n_1} \succ \cdots \succ a_{r,1}, \ldots, a_{r,n_r} \qquad \text{with} \qquad r \geq 1 \qquad \text{and} \qquad n_1 + \cdots + n_r = n.$$

Such rankings represent full rankings with ties, in the sense that all the elements of $[\![n]\!]$ are ranked but for some pairs of elements, the order is not specified. They are the rankings with connected Hasse diagrams. Strict partial rankings are represented by the second row in Table 2.1 for $n = 3$. A partial ranking can also be viewed as an ordered partition $(A_1, \ldots, A_r)$ of $[\![n]\!]$ where the elements of $A_1$ are placed equal first and the elements of $A_r$ are placed equal at $r^{th}$ rank. The subsets $A_i$ are sometimes called "buckets" and the partial rankings *bucket orders*. Of special interest are the top-$k$ rankings, either ordered, of the form $a_1 \succ \cdots \succ a_k \succ$ *the rest*, or unordered, of the form $a_1, \ldots, a_k \succ$ *the rest*, with $1 \leq k \leq n$.

**Incomplete rankings.** Another generalization of full rankings are *incomplete rankings*, of the form

$$a_1 \succ \cdots \succ a_k \qquad \text{with} \qquad 2 \leq k \leq n.$$

They correspond to full rankings restricted to a subset of elements only, and are also called *subset rankings*. Their Hasse diagrams are chains that only involves a subset of elements, the other elements being isolated nodes. Strict incomplete rankings are represented by the third row in Table 2.1 for $n = 3$. Pairwise comparisons, of the form $a_1 \succ a_2$, are an important particular case of incomplete rankings. They are the simplest strict partial orders one could consider. Of course the strict incomplete rankings on $[\![3]\!]$ are pairwise comparisons. For $2 \leq k \leq n$ and distinct elements $a_1, \ldots, a_k \in [\![n]\!]$, we simply denote the incomplete ranking $a_1 \succ \cdots \succ a_k$ by the expression $\pi = a_1 \ldots a_k$. Such an expression is called an *injective word*, its content is the set $c(\pi) = \{a_1, \ldots, a_k\}$ and its length or size is the number $|\pi| = k$. The rank of element $i \in c(\pi)$ in the ranking $\pi$ is denoted by $\pi(i)$. We denote by $\Gamma_n$ the set of all incomplete rankings on $[\![n]\!]$ and by $\Gamma(A) = \{\pi \in \Gamma_n \mid c(\pi) = A\}$ the set of incomplete rankings with content $A$, for any $A \in \mathcal{P}([\![n]\!])$. Notice that $\Gamma([\![n]\!])$ corresponds to $\mathfrak{S}_n$ and that $\Gamma_n = \bigsqcup_{A \in \mathcal{P}([\![n]\!])} \Gamma(A)$.

*Remark* 3 (Incomplete and partial rankings). More generally, one can also consider incomplete and partial rankings, of the form $a_{1,1}, \ldots, a_{1,n_1} \succ \cdots \succ a_{r,1}, \ldots, a_{r,n_r}$ with $r \geq 1$ and $n_1 + \cdots + n_r \leq n$. We point out that the class of such rankings remains strictly included in $\mathfrak{R}_n$: for instance the ranking $1 \succ 2, 3$ and $4, 5 \succ 6$ on $[\![6]\!]$ does not belong to it.

### 2.1.3 General notations

We finish this section with general notations that we use in the thesis.

**Generic.** The cardinality of a finite set $E$ is denoted by $|E|$. The disjoint union of two sets $A$ and $B$ is denoted by $A \sqcup B$ and the strict inclusion of $A$ in $B$ by $A \subsetneq B$. The indicator function of any event $\mathcal{E}$ is denoted by $\mathbb{I}\{\mathcal{E}\}$. For $x \in \mathbb{R} \setminus \{0\}$ we define $\text{sign}(x) = 1$ if $x > 0$ and $-1$ if $x < 0$.

**Functions on finite sets.** For a set $E$ of finite cardinality $|E| < \infty$, we set $\mathcal{P}(E) = \{A \subset E \mid |A| \geq 2\}$ and denote by $L(E) = \{f : E \to \mathbb{R}\}$ the linear space of real-valued functions on $E$. It is equipped with the canonic inner product $\langle f, g \rangle_E = \sum_{x \in E} f(x)g(x)$ and the associated Euclidean norm $\| \cdot \|_E$. The indicator function of a subset $S \subset E$ is denoted by $\mathbb{1}_S$ in general and by $\delta_x$ when $S$ is the singleton $\{x\}$, in which case it is called a Dirac function. The support of a function $f \in L(E)$ is the set $\text{supp}(f) := \{x \in E \mid f(x) \neq 0\}$.

**Probabilistic modeling.** A probability distribution on a finite set is identified with its probability mass function. For a random variable $X$ on a finite set $E$ and a probability distribution $p$ over $E$, the expression $X \sim p$ means that $X$ is drawn from $p$ or equivalently that $p$ is the law of $X$. If $X$ takes its values in a vector space, we denote by $\mathbb{E}[X]$ its expectation and by $\mathbb{E}[X|\mathcal{B}]$ its conditional expectation with respect to the $\sigma$-algebra $\mathcal{B}$.

**Symmetric group.** The set $\mathfrak{S}_n$ of permutations of $[\![n]\!]$ is equipped with the composition operation $\mathfrak{S}_n \times \mathfrak{S}_n \to \mathfrak{S}_n,\ (\sigma, \tau) \mapsto \sigma\tau$ defined by $\sigma\tau(i) = \sigma(\tau(i))$ for all $i \in [\![n]\!]$. This makes it a group, with unity equal to the identity permutation $id$ defined by $id(i) = i$ for all $i \in [\![n]\!]$. The inverse of a permutation $\tau \in \mathfrak{S}_n$ is denoted by $\tau^{-1}$.

**Linear algebra.** Here and throughout the thesis, we consider linear operators on finite-dimensional vector spaces. The composition of two operators $T$ and $T'$ is denoted by $TT'$, and the application on a vector $x$ is denoted by $Tx$ or $T(x)$. The null space of a linear operator $T$ between vector spaces $V$ and $W$ is defined by $\ker T = \{x \in V \mid Tx = 0\}$. The identity operator on a vector space $V$ is usually denoted by $Id_V$ and the identity matrix of size $d$ by $I_d$.

## 2.2 Applications

The statistics literature and more generally all the scientific literature involved with data analysis mostly focus on vector data. This is of course justified by the fact that vectors model the vast majority of observation types. It can therefore be surprising to see how many applications involve ranking data. We propose some general reasons for that.

- **Transitivity of preferences is very natural for human beings.** If a person considers that three elements $a$, $b$ and $c$ are comparable and consciously prefers $a$ to $b$ and $b$ to $c$, then she will necessarily consciously prefer $a$ to $c$. Now of course, her preferences can vary through time or depending on the situation, but for a given context, they will be transitive. It is therefore easy to ask people to express their preferences as partial orders.
- **Many objects are naturally represented by partial orders.** Examples include : lists of distinct elements (such as results of a query, words, genes, tasks), matchings between two sets of same cardinality (represented by permutations), rankings (such as competition results or preference judgments).
- **Partial orders are generic mathematical objects, naturally constructed from cardinal values.** In order to aggregate different pieces of cardinal data (such as results

of experiments or users ratings), transforming them into rankings is a universal way to normalize them and can also provide more robust features.

These factors may explain why ranking data analysis has been the subject of such a large literature, spreading across so many domains, from social choice theory to machine learning, through psychology, economics, statistics, artificial intelligence or operations research. Without being exhaustive, we describe here the main applications that have been considered.

### 2.2.1 Social choice

The analysis of ranking data first appeared in the $18^{th}$ century, with the study of an election system for the French *Académie des Sciences*. In such an election setting, $[\![n]\!]$ is a set of candidates and voter express their opinions under the form of a rankings on $[\![n]\!]$. The goal is then to elect one or several winners. In Borda (1781), Borda showed that the classic *plurality* voting rule[1] suffers from important drawbacks and introduced a new voting rule, now called the *Borda Count*, which satisfies several desirable properties. Condorcet showed however in Condorcet (1785) that in some cases, a candidate who wins against all the others in pairwise duels is not elected by the Borda Count. He therefore introduced an example of voting rule that does not suffer from this drawback (such a voting rule is called today a *Condorcet Method*) but does not satisfy in exchange all the desirable properties of the Borda Count. Though the French *Académie des Sciences* chose to use the Borda Count, this started the still open *Borda-Condorcet debate* (see for instance Risse, 2005), and more generally the study of election systems in social choice theory.

The field took its modern form with the seminal contribution Arrow (1950). The latter defines a general framework to study voting rules through their axiomatic properties. A first result is the well-known "impossibility theorem" (Arrow, 1951): no rule can satisfy simultaneously a predefined set of reasonable properties. Hence, as there is no "good voting rule", each voting rule deserves to be analyzed, with its advantages and drawbacks. This has led the researchers to introduce new voting rules (Copeland, 1951; Young, 1977; Tideman, 2006; Goldsmith et al., 2014), to establish properties for existing ones (Fishburn, 1977; Young and Levenglick, 1978; Barthélémy and Montjardet, 1981), to develop new interpretations and connections (Young, 1988; Saari, 2000; Saari and Merlin, 2000; Kalai, 2002; Daugherty et al., 2009; Lahaie and Shah, 2014), to make empirical comparisons between different voting rules (Mattei, 2011; Popov et al., 2014), or to extend Arrow's framework (Sen, 1970; Gibbard, 1973; Satterthwaite, 1975; Sen, 1977; Balinski and Laraki, 2010; Prasad et al., 2015).

In the early 1990s, Bartholdi et al. (1989, 1992) showed that methods and concepts from computer science were relevant for social choice theory. More and more contributions then developed this approach to end up with the creation of the new field of *Computational Social Choice* with the first COMSOC workshop in 2006. Contributions have for instance studied the computational complexity of winner determination and manipulation (Hemaspaandra et al., 1997; Conitzer et al., 2006; Procaccia and Rosenschein, 2006; Conitzer et al., 2007; Faliszewski et al., 2009; Faliszewski and Procaccia, 2010), the stability of voting procedures under different perturbations (Erdélyi et al., 2011; Lu and Boutilier, 2011b; Procaccia et al., 2012; Caragiannis et al., 2014), introduced new families of voting rules (Conitzer and Sandholm, 2005; Conitzer et al., 2009; Zwicker, 2008; Xia and Conitzer, 2008; Caragiannis et al., 2013; Elkind et al., 2015) and made some connections with machine learning (Dwork et al., 2001; Soufiani et al., 2014b).

[1] Voters only vote for their favorite candidate, the winner is the candidate with the maximal number of votes.

### 2.2.2 Psychometry, statistics and competitions

In the late $19^{th}$ century, psychologists observed variability and imprecision in human judgments (see Fechner, 1860; Titchener, 1901). In an attempt to model this phenomenon and move into the realm of preferences, Thurstone (1927a) introduced the first probabilistic model for ranking data. This seminal contribution laid out the basis of a general approach to psychometry based on paired comparisons (see Guilford, 1954; Torgerson, 1958; Nunnally et al., 1967), which has been applied to analyze value judgments, such as the pleasantness of different colors (Titchener, 1901), seriousness of crimes (Thurstone, 1927b), scenic beauty of forest scenes (Buhyoff and Leuschner, 1978) or seriousness of environmental losses (Brown et al., 2002).

The method of paired comparisons has been developed more generally in the statistics literature by many contributions (see Kendall and Babington Smith, 1940; Babington Smith, 1950; Mosteller, 1951; Bradley and Terry, 1952; Slater, 1961; David, 1963; Bock and Jones, 1968; Tversky and Russo, 1969; Saaty, 1977). Several monographs and surveys give a global view of the subject (see Bradley, 1976; Davidson and Farquhar, 1976; Böckenholt, 2006; Cattelan, 2012).

Among the numerous applications of the developed methods, there has been a particular focus on tournaments and competitions (Kendall, 1955; Buhlmann and Huber, 1963; Jech, 1983; Glickman and Jensen, 2005), with applications to sports Keener (1993); Masarotto and Varin (2012); Cattelan et al. (2013); Barrow et al. (2013), racing Plackett (1975); Henery (1981); Benter (1994); Ali (1998) or chess and gaming (Zermelo, 1929; Elo, 1978; Batchelder and Bershad, 1979; Henery, 1992; Herbrich et al., 2006; Dangauthier et al., 2007; Weng and Lin, 2011; Nikolenko and Sirotkin, 2011).

### 2.2.3 Economic choices

Choice modeling arose in economics in the second half of the $20^{th}$ century as a new approach to analyze the demand. In Marschak (1959), Thurstone's model was given the economic interpretation of a *Random Utility Model (RUM)*, together with Luce's model introduced in Luce (1959). The latter was then fully characterized and used as a conditional logit model, today called the *Multinomial logit model (MNL)*, in McFadden (1974a). This seminal contribution introduced a framework for the economic analysis of choice behavior, leading to a tremendous number of developments (Manski, 1977; McFadden, 1980; Guadagni and Little, 1983; Hausman and McFadden, 1984; Berry et al., 1995; McFadden and Train, 2000; Walker and Ben-Akiva, 2002; Train, 2009; Soufiani et al., 2013a), with a specific focus on travel demand analysis (McFadden, 1974b; Williams, 1977; Ben-Akiva and Lerman, 1985; Ben-Akiva and Bierlaire, 1999).

Discrete choice models have also been extensively studied in operations research, with the typical problem of *assortment optimization*: how to optimize the set of proposed items to a customer in order to maximize the revenue. Contributions have mainly focused on introducing new choice models together with algorithmic procedures to fit and use them (Talluri and Van Ryzin, 2004; Zhang and Cooper, 2005; Natarajan et al., 2009; Li and Huh, 2011; Blanchet et al., 2013; Farias et al., 2013; Gallego et al., 2014; Désir et al., 2015; Berbeglia, 2016).

### 2.2.4 Decision analysis

In the years 1970, more and more researchers tried to help people facing complex decisions by formalizing their possible alternatives and finding their best solutions. As critical cases arise when decisions are evaluated through multiple criteria, which can in addition contradict, this led to the creation of a new field in operations research called Multiple Criteria Decision Making/Analysis (MCDM or MCDA) with a first conference in 1972 (proceedings published in Zeleny and Cochrane, 1973). It has known an important number of developments and applications since

then, with the introduction of different methods and theories such as outranking methods (Roy, 1968, 1991), multi attribute utility theory (Dyer et al., 1992; Wallenius et al., 2008), fuzzy sets (Carlsson and Fullér, 1996) or rough sets (Greco et al., 2001). Several surveys give a good overview (Figueira et al., 2005; Velasquez and Hester, 2013).

Using the article Luce and Tukey (1964) on conjoint measurement, Green and Rao (1971) established conjoint analysis as a subfield of marketing research. The main considered problems are to understand how buyers make complex purchase decisions, to estimate preferences and importances for product features, and to predict buyer behavior, based on the exploitation of ranking data representing preferences over items with multiple features. Many authors have contributed to the development of the field since then, introducing new methods and algorithms (Louviere, 1988; Swait and Louviere, 1993; Lenk et al., 1996; Louviere et al., 2000) or considering new applications (Wittink and Cattin, 1989; Adamowicz et al., 1994; Ryan, 1999; Ryan and Farrar, 2000; Soutar and Turner, 2002). Many surveys have provided overviews of the field (Green and Srinivasan, 1978, 1990; Green et al., 2001; Netzer et al., 2008).

### 2.2.5 Computer systems

Computer systems have revitalized ranking data analysis since the beginning of the $21^{st}$ century. With the abundance of information has come the need for the selection of items best fitted for a certain purpose. Companies have thus designed ranking systems to present items in the best order, either in a search or a recommendation setting.

Search engines are the central application of *information retrieval (IR)*. Though their concept dates back at least to the mid-$20^{th}$ century (Maron and Kuhns, 1960), they have of course attracted more attention with the development of computer systems and the Web (Deerwester et al., 1990; Baeza-Yates and Ribeiro-Neto, 1999). The principle of fitting ranking functions on datasets labeled by human annotators (Fuhr, 1992; Cooper et al., 1992) or click-through data (Joachims, 2002) has then lead to a tremendous number of machine learning-based methods (see Liu, 2009, for a survey), even more boosted by the *Yahoo! Learning to Rank Challenge* (Chapelle and Chang, 2011). If many of these contributions exploit labels as cardinal values and thus do not correspond to ranking data analysis, a significant number of them deal with ordinal comparisons and therefore perform ranking data analysis. A description of the literature is provided in Subsection 2.3.2.

The *metasearch* problem of combining the results of different search engines has also attracted a great deal of attention. Formalizing it as a rank aggregation problem (see Section 2.3) has enabled to apply classic voting rules and introduce new ones with efficiency (Aslam and Montague, 2001; Dwork et al., 2001; Renda and Straccia, 2003; Fagin et al., 2003; Agrawal et al., 2006; Akritidis et al., 2011; Desarkar et al., 2016).

In recommender systems, the goal is to select for each user items from a catalog that they would like. The central problem is thus to infer maximal knowledge, from available data, about the tastes or equivalently the preferences of each user. Research about recommender systems was however boosted by the *Netflix challenge*, between 2006 and 2009, where the problem was formulated as the prediction of the rating that a user would give to an item at a certain date. Impressive advances were made on the problem of matrix completion, in particular about the use of matrix factorization methods and Restricted Boltzmann Machines (Salakhutdinov et al., 2007; Koren et al., 2009). Many important recommending tasks were however left aside, such as how to rank recommendations, how to exploit other feedback than ratings (implicit feedback in particular), how to increase the diversity of recommendations or how to perform cold-start recommendations (see Shani and Gunawardana, 2011, for a general overview). These various problems can be tackled by methods from ranking data analysis and have therefore driven the

literature over the recent years (Jin et al., 2003; Kamishima, 2003; Weimer et al., 2007; Liu and Yang, 2008; Rendle et al., 2009; Baltrunas et al., 2010; Balakrishnan and Chopra, 2012; Volkovs and Zemel, 2012; Sun et al., 2012; Yi et al., 2013; Wang et al., 2014; Kapicioglu et al., 2014; Lee et al., 2014; Lu and Negahban, 2014; Park et al., 2015).

From a global perspective, most of the aforementioned contributions involve *Preference Learning*, and there has been many efforts to gather them in a global framework (see Fürnkranz and Hüllermeier, 2011).

*Remark* 4 (Ranking on graphs and manifolds). With the Success of the PageRank algorithm (Page et al., 1999) used by Google, there has been a dedicated interest in the literature for ranking on graphs and manifolds for search engines (Zhou et al., 2004; Agarwal, 2006; Xu et al., 2011a). Such methods do not however deal with ranking data and this is why we do not take them into account into our description of the field.

### 2.2.6 Crowdsourcing

Crowdsourcing problems have attracted more and more attention in the last few years, facilitated by the *Amazon Mechanical Turk marketplace*. They consist in dividing a global task into several ones, give them to individual "workers" and aggregate the results. Several ranking applications are well tackled by this approach, such as subjective labeling (Bennett et al., 2009; Xu et al., 2011b; Chen et al., 2013; Xu et al., 2014; Stoyanovich et al., 2015; Park et al., 2015), human computation (Pfeiffer et al., 2012; Mao et al., 2013) or peer grading (Walsh, 2014; Raman and Joachims, 2015).

### 2.2.7 Biological data

With the development of microarrays in the recent years, it has become possible to measure the simultaneous level of expression of thousands of genes or proteins in biological experiments. Such data is of interest to better understand the role of each gene or protein under different conditions. The measured levels of expression can however vary a lot between experiments so that normalization is key to efficiently aggregate observations. For that purpose, exploiting the data as lists of genes or proteins ordered by level of expression and applying methods from ranking data analysis has shown to be a simple and powerful approach (Breitling et al., 2004; Geman et al., 2004; Tan et al., 2005; DeConde et al., 2006; Boulesteix and Slawski, 2009; Kolde et al., 2012; Kim et al., 2014; Jiao and Vert, 2015). With its democratization in the bioinformatics literature, ranking data analysis has found other applications, such as nanotoxicology (Patel et al., 2013) or brain data analysis (Shadi et al., 2015).

### 2.2.8 Mathematical applications

Rankings and permutations arise in many mathematical problems so that methods to analyze them have naturally been found an interest. In machine learning and statistics, they have been applied for instance to multi-class classification (Friedman, 1996; Hastie and Tibshirani, 1998; Huang et al., 2006), feature selection (Wu et al., 2009; He and Yu, 2010; Prati, 2012; Dittman et al., 2013), metric learning (Schultz and Joachims, 2004; Chechik et al., 2010; Zheng et al., 2013), reinforcement learning (Akrour et al., 2011; Cheng et al., 2011; Wilson et al., 2012), algorithms benchmarking (Hornik and Meyer, 2007; Eugster et al., 2014; Mersmann et al., 2015), or the analysis of nonparametric statistical tests (Haunsperger and Saari, 1991; Diaconis et al., 2001; Bargagliotti, 2009; Bargagliotti and Saari, 2010). In combinatorial optimization, though the classic *Quadratic Assignment Problem* has been extensively studied through a quadratic

programming approach (Pardalos et al., 1994), new insights and methods have come from permutation analysis (Barvinok and Vershik, 1988; Kondor, 2010). At last, permutation analysis is naturally involved the study of symmetries (Viana, 2006; Jiang et al., 2014).

### 2.2.9 Diverse

Ranking data and permutations analysis have found many other applications such as multi-object and identity tracking (Kondor et al., 2007; Jiang et al., 2011a), image segmentation (Yu, 2009; Maire, 2010; Yu, 2012), photo sequencing (Basha et al., 2012), image association (Pachauri et al., 2012, 2014) or geometric model fitting (Wong et al., 2013) in image processing, but also seriation (Fogel et al., 2013; Lim and Wright, 2014) link prediction in complex networks (Pujari and Kanawati, 2012; Tabourier et al., 2014) or data coding (Barg and Mazumdar, 2010; Helmi et al., 2012; Wang et al., 2013; Farnoud et al., 2014).

## 2.3 Classic problems in ranking data analysis

In this section we formalize the main problems that have been considered in ranking data analysis. We differentiate between problems on rankings of elements without features and problems on rankings of elements with features or with context. Though the contributions of this thesis apply more directly to the former setting, they could be applied to the latter (examples of directions are described in Chapter 7). This is why we describe the main problems considered in the literature for both settings.

### 2.3.1 Rankings of elements without features

In its widest generality, a dataset of rankings is a collection of $N \geq 1$ rankings $(\pi^{(1)}, \dots, \pi^{(N)}) \in \mathfrak{R}_n^N$. Though it is not always necessary, it is natural to model it as a collection of random rankings $\mathcal{D}_N = (\Pi^{(1)}, \dots, \Pi^{(N)}) \in \mathfrak{R}_n^N$ drawn IID from a probability distribution $\mu$ over $\mathfrak{R}_n$. The task can then be to describe, summarize or visualize the dataset or to infer some target part of the probability distribution $\mu$. The former corresponds to descriptive statistics or data visualization, and the latter corresponds to inferential statistics or unsupervised learning. Most of the usual tasks on vector data can be considered on ranking data but they usually require the definition of specific concepts and the design of dedicated methods. Other tasks are also particular to ranking data. In addition, while most of the tasks can be considered for any subclass of rankings, they can be much more interesting and/or challenging for some subclasses compared to others. The main problems are described below.

**Rank aggregation.** Rank aggregation was the first problem to be considered on ranking data and has been the most widely studied one in the literature. The goal is to find one full ranking $\sigma^* \in \mathfrak{S}_n$ that best "represents" the data. While traditional vector data is naturally represented by its mean, there is no equivalent concept for ranking data and there are many ways to state the problem of rank aggregation formally. One approach widely considered in the literature is to look for *consensus rankings* (Kemeny, 1959). Given a dissimilarity measure $\Delta$ on $\mathfrak{R}_n$, they are defined in the following ways.

- **Summary setting.** A consensus ranking for a dataset $\mathcal{D}_N = (\Pi^{(1)}, \dots, \Pi^{(N)}) \in \mathfrak{R}_n^N$ is a permutation $\sigma^* \in \mathfrak{S}_n$ solution of

$$\min_{\sigma \in \mathfrak{S}_n} \sum_{t=1}^{N} \Delta\left(\sigma, \Pi^{(t)}\right). \tag{2.1}$$

- **Inference setting.** A consensus ranking for a probability distribution $\mu$ over $\mathfrak{R}_n$ is a permutation $\sigma^* \in \mathfrak{S}_n$ solution of

$$\min_{\sigma \in \mathfrak{S}_n} \sum_{\pi \in \mathfrak{R}_n} \Delta(\sigma, \pi)\mu(\pi). \tag{2.2}$$

Though less common, Inference setting (2.2) is considered for instance in Prasad et al. (2015) or Rajkumar and Agarwal (2014). For two full rankings $\sigma, \sigma' \in \mathfrak{S}_n$, $\Delta(\sigma, \sigma')$ is typically equal to a distance on $\mathfrak{S}_n$ (see Subsection 2.4.4 for examples), and for a pairwise comparison $i \succ j$, one typically takes $\Delta(\sigma, i \succ j) = \mathbb{I}\{\sigma(i) > \sigma(j)\}$.

The most widely considered case is arguably the Summary setting (2.1) for full rankings with dissimilarity measure equal to the Kendall's tau distance on $\mathfrak{S}_n$ (refer to Subsection 2.4.4 for the definition). It has indeed been shown that the voting rule that maps a dataset to its associated consensus(es) for this setting, called *Kemeny's rule*, is the unique rule that satisfies some desirable axiomatic properties (Young and Levenglick, 1978) and that the consensus(es) are the maximum likelihood estimator(s) for the Mallows model (Young, 1988), refer to Subsection 2.5 for the definition. *Kemeny rank aggregation*, the problem of finding a Kemeny consensus, is however NP-hard (Bartholdi et al., 1989; Dwork et al., 2001; Hudry, 2008). Many contributions have thus tackled this complexity by establishing theoretical guarantees for existing procedures (see Diaconis and Graham, 1977; Saari and Merlin, 2000; Coppersmith et al., 2006; Freund and Williamson, 2015), complexity bounds for exact recovery under hypothesis on the dataset (see Davenport and Kalagnanam, 2004; Conitzer et al., 2006; Brandt et al., 2015) or if some quantity is known on the dataset (see for instance Betzler et al., 2008, 2009; Karpinski and Schudy, 2010; Fernau et al., 2010; Cornaz et al., 2013; Betzler et al., 2014). Many others have introduced alternative rank aggregation procedures (in addition of the voting rules already mentioned in Subsection 2.2.1), seen as approximations of Kemeny's rule (Van Zuylen and Williamson, 2007; Kenyon-Mathieu and Schudy, 2007; Ailon et al., 2008; Van Zuylen and Williamson, 2009) - with comparisons on numerical experiments (Schalekamp and van Zuylen, 2009; Ali and Meila, 2012) - or not necessarily (Blin et al., 2011; Niu et al., 2013; Aledo et al., 2013; Deng et al., 2014; Lorena et al., 2014; Volkovs and Zemel, 2014; Bedo and Ong, 2014). Some contributions have also considered the rank aggregation problem with other distances (Chin et al., 2004; Bachmaier et al., 2013; Farnoud and Milenkovic, 2014).

With large-scale modern applications (see Section 2.2), another important literature has been devoted to aggregation from pairwise comparisons (Braverman and Mossel, 2008; Gleich and Lim, 2011; Jiang et al., 2011b; Yu, 2012; Negahban et al., 2012; Wauthier et al., 2013; Rajkumar and Agarwal, 2014; Xu et al., 2014; Cucuringu, 2015; Shah and Wainwright, 2015) or partial rankings (Fagin et al., 2004; Ailon, 2010; Brandenburg et al., 2012; Procaccia and Shah, 2015) or both (Ammar and Shah, 2012).

**Partial rank aggregation.** Aggregating rankings into a full ranking can sometimes be too hard and unnecessary. This is why the literature has also considered the problem of aggregating rankings into a top-$k$ ranking or a more general partial ranking. The latter states as (2.1) except that one looks for a partial ranking $\pi^*$ instead of a full ranking $\sigma^*$ (Gionis et al., 2006; Feng et al., 2008; Kenkre et al., 2011). A "parametric" ranking model is usually assumed in the former (see Subsection 2.5.1 for examples). As classic parametric models have a natural associated ordering of the elements of $[\![n]\!]$, the problem is then to construct an estimator of the top-$k$ elements from the dataset and its accuracy is measured by the probability that it is correct under the model. An extension of the Mallows model is used in Procaccia et al. (2012) while the Plackett-Luce model is used in Chen and Suh (2015); Jang et al. (2016). A "nonparametric" setting (see Subsection

2.5 for a definition) is also considered in Rajkumar et al. (2015) where the considered problem is to recover the winners (for some criterion) of the true ranking model, under several "regularity assumptions".

**Estimation.** Estimating the probability distribution that underlies the data generation is the central task of statistical ranking data analysis. Any introduction of a new model thus usually comes with an associated inference procedure, that we describe in Section 2.5. We simply point out that most of the literature do not seek to estimate the probability distribution $\mu$ over $\mathfrak{R}_n$. The latter is indeed usually decomposed as a product between a ranking model and an observation design and the goal is to estimate the ranking model (see Section 3.1 for the proper definitions).

While many contributions have extended existing ranking models and estimation methods or introduced new ones, very few have developed a statistical theory for ranking data analysis. This is of course due to its many challenges and specificities (see Section 2.4). We however point out several contributions, that have introduced notions of "confidence interval" (Patil and Taillie, 2004; Hall and Miller, 2010; Volkovs and Zemel, 2014), focused on bounds for existing methods (Maystre and Grossglauser, 2015b; Khetan and Oh, 2016) or established minimax-optimality results (Hajek et al., 2014; Shah et al., 2015b).

**Clustering.** Clustering is a natural problem in ranking data analysis, especially in the numerous applications where the data represent the preferences of a users population (see Section 2.2). The dataset is assumed to be of the form $\mathcal{D}_N = ((\Pi^{(1,1)}, \dots, \Pi^{(1,N_1)}), \dots, (\Pi^{(m,1)}, \dots, \Pi^{(m,N_m)})) \in \mathfrak{R}_n^N$, where $m$ is the number of users, $(\Pi^{(j,1)}, \dots, \Pi^{(j,N_j)})$ is the collection of preferences expressed by user $j$ for each $j \in \{1, \dots, m\}$ and $N = N_1 + \cdots + N_m$ is the total number of observations. Two tasks are involved in the clustering problem:

1. Divide the dataset $\mathcal{D}_N$ into clusters.

2. Affect each (potentially new) user $j$ with preferences $(\Pi^{(j,1)}, \dots, \Pi^{(j,N_1)})$ to a cluster.

The vast majority of the literature tackle this problem via the estimation of a mixture of ranking models, such as extensions of the Mallows model (Murphy and Martin, 2003; Meila and Chen, 2010; Lee and Yu, 2012; Awasthi et al., 2014; Jacques and Biernacki, 2014; Chierichetti et al., 2015; Ding et al., 2015a), of the Plackett-Luce model (Busse et al., 2007; Gormley and Murphy, 2008, 2009; Oh and Shah, 2014; Tran and Venkatesh, 2014; Wu et al., 2015; Oh et al., 2015; Mollica and Tardella, 2015) or of the Thurstone model (Abbasnejad et al., 2013). A non parametric approach is also proposed in Clémençon et al. (2011).

**Ranking prediction and collaborative ranking.** Ranking prediction has mostly been considered in the setting of *collaborative ranking*. Mainly applied to a recommendation setting with $m$ users, the goal is to predict a full ranking on $[\![n]\!]$ for each user, based on their feedback on some elements of $[\![n]\!]$. Feedback can be given as ratings or ordinal preferences. The latter case can be formally stated in a supervised learning setting: feedback from user $j$ is modeled by a random ranking $\Pi^j$ drawn from probability distribution $\mu_j$ over $\mathfrak{R}_n$ and the objective is to minimize the theoretical risk defined for a tuple of $m$ full rankings $(\sigma_1, \dots, \sigma_m) \in \mathfrak{S}_n^m$ by

$$\mathcal{R}(\sigma_1, \dots, \sigma_m) := \sum_{j=1}^{m} \mathbb{E}\left[l\left(\sigma_j, \Pi^j\right)\right] = \sum_{j=1}^{m} \sum_{\pi \in \mathfrak{R}_n} l(\sigma_j, \pi)\, \mu_j(\pi),$$

where $l : \mathfrak{S}_n \times \mathfrak{R}_n \to \mathbb{R}^+$ is a given loss function. Of course the true distributions $\mu_j$ are not known and one considers the empirical version of the risk defined for $(\sigma_1, \dots, \sigma_m) \in \mathfrak{S}_n^m$ by

$$\mathcal{R}_N(\sigma_1, \dots, \sigma_m) = \sum_{j=1}^{m} \sum_{t=1}^{N_j} l\left(\sigma^{(m)}, \Pi^{(j,t)}\right),$$

where the $\Pi^{(j,t)}$'s for $t = 1, \dots, N_j$ are drawn IID from $\mu_j$ for each user $j$. For arbitrary probability distributions $\mu_j$, this problem would be equivalent to $m$ independent ones. Empirical data shows however that they have some similarities and that it is more efficient solve the problem at once with some regularization scheme. Contributions in the literature have thus applied methods from machine learning (Liu and Yang, 2008; Rendle et al., 2009; Volkovs and Zemel, 2012; Kuang et al., 2016), especially matrix factorization techniques (Weimer et al., 2007; Balakrishnan and Chopra, 2012; Yi et al., 2013; Wang et al., 2014; Kapicioglu et al., 2014; Lu and Negahban, 2014; Lee et al., 2014; Park et al., 2015; Barjasteh et al., 2015; Oh et al., 2015).

**Active ranking and preference elicitation.** Active learning has been applied to ranking data analysis in mainly two problems. *Active ranking* consists in a rank aggregation problem (full or partial), usually from pairwise comparisons, where the ranker can choose which pairs to observe. Formally, for each chosen pair $\{a_t, b_t\} \subset [\![n]\!]$, he observes the ranking $\Pi^{(t)} \in \{a_t \succ b_t, a_t \prec b_t\} =: \Gamma(\{a_t, b_t\})$ drawn from a probability distribution $P_{\{a_t,b_t\}}$ over $\Gamma(\{a_t, b_t\})$. The performance of a method is measured by the number of queries required to recover the target (exactly or approximately). Contributions in the literature have introduced several methods depending on the considered target and a possible assumption on the $P_{\{a,b\}}$'s (Jamieson and Nowak, 2011; Ailon, 2012; Eriksson, 2013; Busa-Fekete et al., 2013, 2014b).

The second problem is called *Preference elicitation*. Assuming a parametric model on the data, the goal is to choose pairs to observe in order to best approximate the parameters in a minimum number of queries. Contributions in the literature have introduced observation schemes for the Thurstone-Mosteller model or its generalizations (Brochu et al., 2008; Guo and Sanner, 2010; Guo et al., 2010; Pfeiffer et al., 2012; Houlsby et al., 2012; Soufiani et al., 2013a) or the Mallows model (Busa-Fekete et al., 2014a).

**On-line permutation learning.** On-line permutation learning is the following problem: at each step $t = 1, \dots, N$, a learner predicts a permutation $\Sigma^{(t)}$, usually generated by a stochastic algorithm, suffers a loss $l_t(\Sigma^{(t)})$, and is revealed some feedback. The performance of the learner is measured by the difference between its expected total over the $N$ rounds and the total loss of the optimal static permutation in hindsight:

$$\mathcal{R}\left(\Sigma^{(1)}, \dots, \Sigma^{(N)}\right) = \sum_{t=1}^{N} \mathbb{E}\left[l_t\left(\Sigma^{(t)}\right)\right] - \min_{\sigma \in \mathfrak{S}_n} \sum_{t=1}^{N} l_t(\sigma).$$

In Yasutake et al. (2012), a permutation $\sigma^{(t)} \in \mathfrak{S}_n$ is revealed to the learner at each step and the latter suffers the loss $l_t(\sigma) = d_{KT}(\sigma, \sigma^{(t)})$, where $d_{KT}$ is the Kendall's tau distance on $\mathfrak{S}_n$ (see Subsection 2.4.4 for the definition). On-line permutation learning it this case can then be seen as a rank aggregation problem where the ranker is presented the rankings of a collection $(\sigma^{(1)}, \dots, \sigma^{(N)}) \in \mathfrak{R}_n^N$ one-by-one and asked at each step to take a guess on the consensus for the full collection. Other losses and types of feedback have been considered in the literature for which the contributions have each introduced specific methods (Helmbold and Warmuth, 2009; Yasutake et al., 2012; Ailon, 2014; Ailon et al., 2014; Chaudhuri and Tewari, 2015).

**Visualization.** Visualizing ranking data is of course very useful for its analysis but it represents a great challenge, because of its combinatorial nature and high dimensionality. Several contributions in the literature have thus proposed visualization methods, mainly based on dimensionality reduction (Yu and Chan, 2001; Ukkonen, 2007; Kidwell et al., 2008; Sun et al., 2010) or graphical techniques (Shi et al., 2012; Gratzl et al., 2013; Behrisch et al., 2013; Lei et al., 2016).

### 2.3.2 Rankings of elements with features or with context

In this subsection we consider the main problems of ranking data analysis with features. Because they do not constitute direct applications of our results, we only describe some of the problems encountered in the abundant associated literature.

**Label ranking.** Let $\mathcal{X}$ represent an input set (typically $\mathcal{X} = \mathbb{R}^d$) and let $[\![n]\!]$ represent a set of "labels". Label ranking consists in learning a function $f$ that predicts a ranking $\pi_x \in \mathfrak{R}_n$ on $[\![n]\!]$ for each input $x \in \mathcal{X}$, from the observations of IID samples of a random couple $(X, \Pi) \in \mathcal{X} \times \mathfrak{R}_n$. The performance of the function $f$ is evaluated by its risk

$$\mathcal{R}(f) = \mathbb{E}[l(f(X), \Pi)],$$

where $l : \mathfrak{R}_n \times \mathfrak{R}_n \to \mathbb{R}^+$ is a loss function. This setting generalizes multi-class and multi-label classification: in the former, observations are top-1 rankings (the preferred element being the class of the input) and in the latter, observations are of the form $a_1, \ldots, a_k \succ b_1, \ldots, b_{n-k}$ (the $a_i$'s corresponding to the labels equal to 1 for the given input and the $b_i$'s to the labels equal to 0). Methods introduced in the literature rely for instance on kernels (Elisseeff and Weston, 2001; Chu and Ghahramani, 2005b), decomposition on pairwise comparisons (Fürnkranz, 2002; Hüllermeier et al., 2008; Destercke, 2013), boosting (Dekel et al., 2003), local regularity (Brinker and Hüllermeier, 2007; Cheng et al., 2009, 2010) or on-line procedures (Crammer and Singer, 2003; Shalev-Shwartz and Singer, 2007; Grbovic et al., 2013).

**Ordinal regression, instance/object ranking and learning to rank.** In this paragraph we consider an infinite set $\mathcal{X}$ of elements to be ranked, each characterized by its features, typically $\mathcal{X} = \mathbb{R}^d$. A scoring function $f : \mathcal{X} \to \mathbb{R}$ induces a strict partial order on $\mathcal{X}$ through $x \succ x'$ if and only if $f(x) > f(x')$. From a general point of view, ordinal regression, instance/object ranking and learning to rank all consist in learning a scoring function $f : \mathcal{X} \to \mathbb{R}$ from observations of potentially different types:

- **Pointwise feedback.** $(x, r)$ with $x \in \mathcal{X}$ an element and $r \in \mathbb{R}$ a rating or a relevance level.
- **Pairwise feedback.** $x \succ x'$ with $x, x' \in \mathcal{X}$ two distinct elements.
- **Listwise feedback.** $x_1 \succ \cdots \succ x_k$ with $x_1, \ldots, x_k \in \mathcal{X}$ $k$ distinct elements and $k \geq 2$.

Many methods in the literature transform feedback from one type to another and many loss functions have been introduced to take each setting into account (see Liu, 2009). One can nevertheless globally differentiate between pointwise methods (Crammer and Singer, 2001; Shashua and Levin, 2002; Chu and Ghahramani, 2005a; Cossock and Zhang, 2006; Li et al., 2007), pairwise methods (Herbrich et al., 1999; Cohen et al., 1999; Freund et al., 2003; Burges et al., 2005; Pahikkala et al., 2007; Ailon and Mohri, 2010) and listwise methods (Burges et al., 2006; Cao et al., 2007; Xu and Li, 2007; Yue et al., 2007; Taylor et al., 2008; Xia et al., 2008; Cossock and Zhang, 2008; Pareek and Ravikumar, 2014). Theoretical guarantees have also been studied

(Clémençon and Vayatis, 2007; Clémençon et al., 2008; Ravikumar et al., 2011; Duchi et al., 2013). Refer to Liu (2009); Busa-Fekete et al. (2012); Tax et al. (2015) for surveys and comparisons. Because they deal with ranking data, pairwise and listwise methods can be considered as part of ranking data analysis.

## 2.4 Specificities and challenges of ranking data analysis

Many of the problems described previously can be stated in a traditional statistics or machine learning setting. Solving them for ranking data needs however to deal with specific challenges. We illustrate them on the estimation task in the case of full rankings. Let then $p$ be a probability distribution on the symmetric group $\mathfrak{S}_n$. We assume to observe a dataset $\mathcal{D}_N = (\Sigma^{(1)}, \dots, \Sigma^{(N)}) \in \mathfrak{S}_n^N$ of $N$ IID samples from $p$ and the goal is to recover $p$.

*Example* 5 (German dataset). As a running example, we will consider a real dataset obtained from Croon (1989) and studied for example in Diaconis and Sturmfels (1998) or Yao and Böckenholt (1999). After the fall of the Berlin wall, a survey of German citizens was conducted where they were asked to rank four political goals:

1. Maintain order
2. Give people more say in government
3. Fight rising prices
4. Protect freedom of speech

This dataset contains the answers of 2,262 respondents, summarized in the following table.

| Ranking | Answers | Ranking | Answers |
|---|---|---|---|
| 1234 | 137 | 3124 | 330 |
| 1243 | 29 | 3142 | 294 |
| 1324 | 309 | 3214 | 117 |
| 1342 | 255 | 3241 | 69 |
| 1423 | 52 | 3412 | 70 |
| 1432 | 93 | 3421 | 34 |
| 2134 | 48 | 4123 | 21 |
| 2143 | 23 | 4132 | 30 |
| 2314 | 61 | 4213 | 29 |
| 2341 | 55 | 4231 | 52 |
| 2413 | 33 | 4312 | 35 |
| 2431 | 39 | 4321 | 27 |

Throughout the thesis, this dataset is called the *German dataset*.

### 2.4.1 Difference with multivariate analysis

A permutation $\sigma \in \mathfrak{S}_n$ can be represented by the vector $(\sigma(1), \dots, \sigma(n)) \in \mathbb{R}^n$. One can therefore see each sample $\Sigma^{(i)}$ as the random vector $(\Sigma^{(i)}(1), \dots, \Sigma^{(i)}(n))$ and estimate $p$ using techniques from multivariate analysis. This is however infringed by two critical differences.

- The variables $\Sigma(1), \dots, \Sigma(n)$ of a random permutation $\Sigma$ are far from being independent. More specifically, their dependence structure is of combinatorial nature: they must take their values in $[\![n]\!]$ and these values must be different. It is not captured well by classic techniques from multivariate analysis.

- The average of two permutation vectors $(\sigma(1), \ldots, \sigma(n))$ and $(\sigma'(1), \ldots, \sigma'(n))$ is usually not a permutation vector. The Law of Large Numbers and Central Limit Theorems can therefore not be applied.

### 2.4.2 Exploding cardinality of $\mathfrak{S}_n$

From another point of view, the symmetric group $\mathfrak{S}_n$ is a finite set. The distribution $p$ can thus simply be estimated by the histogram of the frequencies of observation, formally by the following empirical estimator

$$\widehat{p}_N = \frac{1}{N} \sum_{t=1}^{N} \delta_{\Sigma^{(t)}},$$

where we recall that $\delta_\sigma$ is the Dirac function on $\sigma$. Problem is that $\mathfrak{S}_n$ has exploding cardinality: $|\mathfrak{S}_n| = n!$. This approach thus becomes irrelevant very quickly when $n$ increases. For instance in the *sushi dataset* from Kamishima (2003) with 5000 permutations of $[\![10]\!]$, the maximal number of observations for one permutation is 3. The exploding cardinality of $\mathfrak{S}_n$ also brings of course a daunting computational challenge, an omnipresent burden of ranking data analysis.

### 2.4.3 Difference with probability density function estimation

If $\mathfrak{S}_n$ is too big to be treated as a finite set, one could treat it as an infinite set and apply methods from probability density function estimation. Figure 2.1 gives an example of a kernel-based estimator for the German dataset. There is however no natural way to order the permutations along an axis and two different orderings can lead to very different results. Figure 2.2 displays the smoothed estimator obtained on the German dataset with the same kernel but after a reordering of the permutations along the axis. The comparison between the two smoothed estimators is provided in Figure 2.3. This example illustrate how arbitrary such an approach is.

### 2.4.4 Absence of a canonical structure

In vector data analysis, statistical estimation usually relies on an hypothesis about the relationship between the probability distribution and some geometrical or topological structure of the space. This hypothesis can take the form of an explicit formula in a parametric approach (e.g. a Gaussian model on $\mathbb{R}^d$ is defined with respect to the Euclidean distance) or a regularity assumption in a non-parametric approach (e.g. a differentiability assumption on $\mathbb{R}^d$ exploits the infinitesimal structure of $\mathbb{R}^d$). The same principle applies to $\mathfrak{S}_n$, but the problem is that there are many possible geometrical or topological structures and none is canonical.

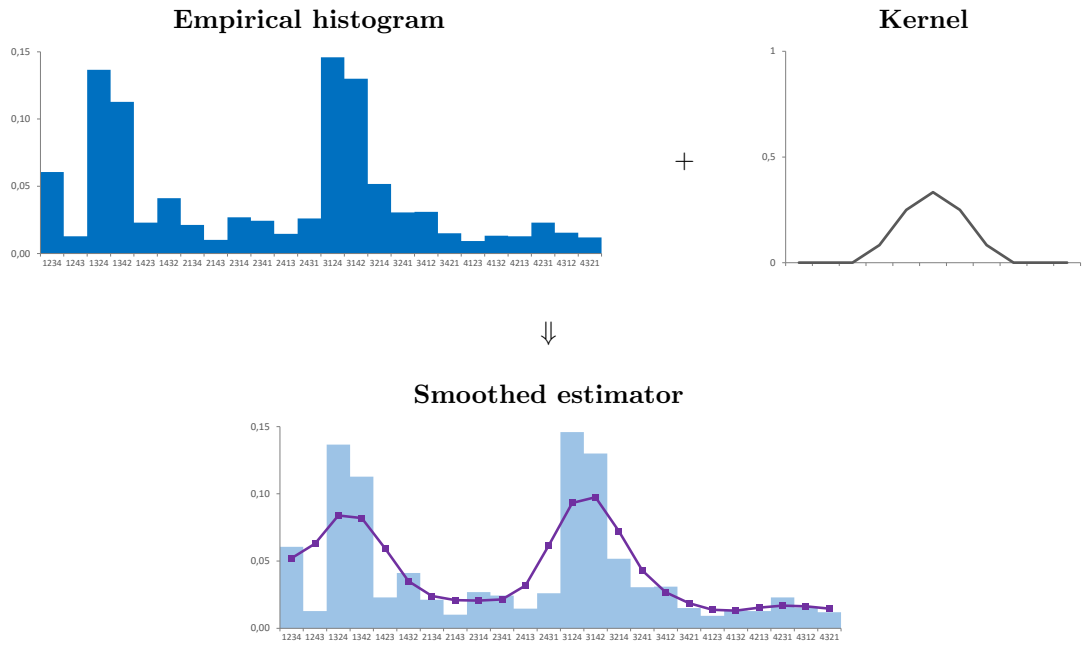



Figure 2.1: Kernel-based estimation on the German dataset

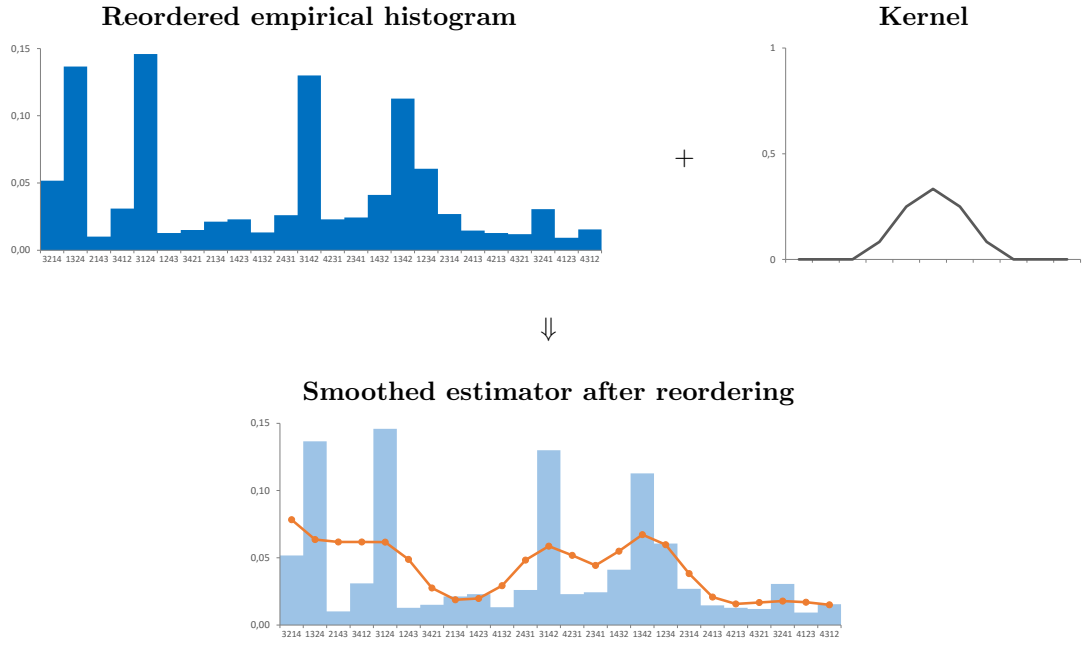



Figure 2.2: Kernel-based estimation on the German dataset

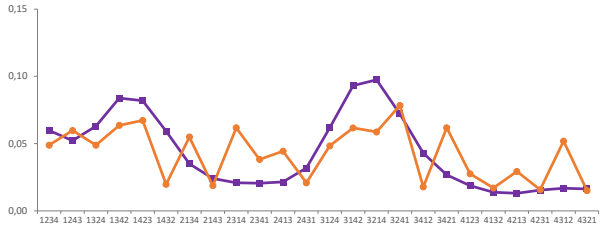



Figure 2.3: Comparison of the two smoothed estimators on the German dataset

**Distances.** Many distances can be defined on $\mathfrak{S}_n$, each one having its own interest. Some classic examples are, for $\sigma, \sigma' \in \mathfrak{S}_n$:

$$\begin{aligned}
\text{Kendall's tau distance} \qquad & d(\sigma,\sigma') = \sum_{1\le i<j\le n} \mathbb{I}\{(\sigma(j)-\sigma(i))(\sigma'(j)-\sigma'(j)) < 0\} \\
\text{Spearman's footrule } (l^1 \text{ distance}) \qquad & d(\sigma,\sigma') = \sum_{i=1}^{n} |\sigma(i)-\sigma'(i)| \\
\text{Spearman's rho } (l^2 \text{ distance}) \qquad & d(\sigma,\sigma') = \sqrt{\sum_{i=1}^{n} (\sigma(i)-\sigma'(i))^2} \\
\text{Hamming distance } (l^0 \text{ distance}) \qquad & d(\sigma,\sigma') = \sum_{i=1}^{n} \mathbb{I}\{\sigma(i) \neq \sigma'(i)\}
\end{aligned}$$

Refer to Deza and Huang (1998) for more examples.

**Graph structures.** The symmetric group can also be given a graph structure: each permutation is seen as a vertex and two permutations are linked if they satisfy some relationship. The usual approach is to consider a *Cayley graph*, where permutations $\sigma$ and $\sigma'$ are linked if and only if $\sigma^{-1}\sigma'$ belongs to a given generating subset $S \subset \mathfrak{S}_n$. Among the many possibilities, examples are:

$$\begin{aligned}
\text{All transpositions :} \qquad & S = \{(i\ j) \mid 1 \le i < j \le n\} \\
\text{Adjacent transpositions :} \qquad & S = \{(i\ i+1) \mid 1 \le i \le n-1\} \\
\text{Star graph :} \qquad & S = \{(i\ n) \mid 1 \le i \le n-1\}
\end{aligned}$$

**Embeddings.** Another approach is to embed $\mathfrak{S}_n$ in a space $V$ equipped with a particular structure, typically a Euclidean space, and transfer this structure to $\mathfrak{S}_n$. Here again, the possibilities are numerous, classic examples being:

$$\begin{aligned}
\text{Embedding as a permutation vector:} \qquad & \mathfrak{S}_n \to \mathbb{R}^n, \qquad \sigma \mapsto (\sigma(1),\dots,\sigma(n)) \\
\text{Embedding as a permutation matrix:} \qquad & \mathfrak{S}_n \to \mathbb{R}^{n\times n}, \qquad \sigma \mapsto P_\sigma = [\mathbb{I}\{\sigma(i)=j\}]_{1\le i,j\le n} \\
\text{Embedding as an acyclic graph:} \qquad & \mathfrak{S}_n \to \mathbb{R}^{n(n-1)/2}, \qquad \sigma \mapsto (\text{sign}(\sigma(j)-\sigma(i)))_{1\le i<j\le n}
\end{aligned}$$

### 2.4.5 Interest for an interpretation

Not being specific to the domain, interpretable models are particularly valued in ranking data analysis. This is surely due to two aspects mentioned previously.

- Ranking data is very natural for human beings and represent in many situations preferences expressed by individuals.
- Many statistical problems enter an unsupervised setting, thus where the ground truth is unknown.

As a consequence, the exploitation of one structure or another on $\mathfrak{S}_n$ should be motivated by an explicit reason.

## 2.5 Models

We now describe the main models that have been introduced in the literature to analyze ranking data. We differentiate between parametric and nonparametric models. This distinction may be surprising because a parametric model in statistics is usually defined as a a collection of probability distributions that can be injected into a finite-dimensional vector space. As $\mathfrak{R}_n$ is a finite set, any probabilistic model on ranking data is parametric under this definition. The distinction we make here rather relies on the associated statistical modeling approaches. In parametric modeling, the model is defined by an explicit formula with parameters and its complexity is characterized by the number of parameters. In nonparametric modeling the model does not always have an explicit expression and more importantly its complexity is characterized by a regularity assumption. This general distinction in statistics applies to ranking data analysis. Another difference is that parametric models on ranking data are usually motivated by a psychological interpretation whereas nonparametric models by a mathematical interpretation.

### 2.5.1 Parametric models

Parametric modeling constitutes the vast majority of the ranking data analysis literature. Perhaps surprisingly however, the contributions can be divided for the major part into three category, each related to a seminal model. We also describe some other parametric models that have been introduced in the literature at the end of this subsection.

**The Thurstone-Mosteller-RUM model.** The first probabilistic model on ranking data was introduced in Thurstone (1927a). Therefore called the *Thurstone model*, it states in its most general form as follows. Given a random vector $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ with probability density function $f : \mathbb{R}^n \to \mathbb{R}^+$, the probability of a full ranking $\sigma = \sigma_1 \dots \sigma_n \in \mathfrak{S}_n$ is equal to

$$p(\sigma) = \mathbb{P}[X_{\sigma_1} > \dots > X_{\sigma_n}] = \int_{x_1 > \dots > x_n} f(x_1, \dots, x_n) dx_1 \dots dx_n.$$

Without any further assumption, this formulation is of course useless because there are much more probability density functions $f$ on $\mathbb{R}^n$ than probability distributions $p$ on $\mathfrak{S}_n$. Two main assumptions are usually made:

1. The $X_i$'s are independent, each with density $f_i : \mathbb{R} \to \mathbb{R}^+$, or equivalently $f$ is the product of the $f_i$'s: $f(x_1, \dots, x_n) = f_1(x_1) \dots f_n(x_n)$ for all $(x_1, \dots, x_n) \in \mathbb{R}^n$. In addition the $f_i$'s are usually assumed to be of the same parametric form with different parameters.

2. $f$ is a multivariate normal distribution of mean vector $\mu \in \mathbb{R}^n$ and covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{n \times n}$.

While a specific case of the first assumption leads to the Plackett-Luce model (see next paragraph), most of the literature that uses the Thurstone model considers the second assumption. Many contributions even consider the simplest case where $\mathbf{\Sigma} = sI_n$, with $s \in \mathbb{R}$ and $I_n$ the identity matrix of size $n$, that is to say where the $X_i$'s are independent and all with same standard deviation $s$. This setting is called case V in Thurstone (1927a) and a least squares regression method was introduced in Mosteller (1951) to fit it on the data. It is usually referred to as the Thurstone-Mosteller in the literature on paired comparisons.

The initial goal of Thurstone was to model the psychological mechanism that leads to people's comparisons. Under case V assumption the interpretation is the following: when a person is presented elements of the set $[\![n]\!]$, she has an unconscious rating $\mu_i$ for each element $i \in [\![n]\!]$ but makes comparisons with the noisy versions $X_i = \mu_i + \epsilon$ where $\epsilon$ is a Gaussian centered noise of standard deviation $s$. The Thurstone model was then given an economics interpretation in Marschak (1959) and called the *Random Utility Model (RUM)*. Many contributions have then extended it (Takane, 1987; Böckenholt, 1992; Maydeu-Olivares, 1999; Yu and Chan, 2001; Walker and Ben-Akiva, 2002; Böckenholt, 2006; Herbrich et al., 2006) or studied and introduced inference methods (Yao and Böckenholt, 1999; Tsai and Yao, 2000; Alberto Maydeu-Olivares and Hernández, 2007; Weng and Lin, 2011; Soufiani et al., 2014a).

**The Bradley-Terry-Luce-Plackett-MNL model.** The *Bradley-Terry model*, introduced in Bradley and Terry (1952), is certainly with the Thurstone-Mosteller model the most widely used model for exploiting pairwise comparisons. Given positive parameters $w_1, \dots, w_n$, it models a pairwise probability as

$$\mathbb{P}[a \succ b] = \frac{w_a}{w_a + w_b},$$

and is classically fitted through maximum likelihood estimation. Interestingly, the same model was already considered in Zermelo (1929) (but it was not known from the Anglo-Saxon literature until the 1960s) and was introduced independently in Ford (1957). In Luce (1959), the author generalizes the model and give it at the same time a new meaning. He shows that it is derived from any choice model (see Section 2.3) that satisfies the *Choice axiom*: for $A, B \in \mathcal{P}([\![n]\!])$ with $A \subset B$ and $a \in A$, the probability of choosing $a$ among $A$ is equal to the probability of choosing $a$ among $B$ conditional on $A$ having been chosen. This axiom leads more generally to a probabilistic model over full rankings, given for $\sigma = \sigma_1 \dots \sigma_n \in \mathfrak{S}_n$ by

$$p(\sigma) = \prod_{i=1}^{n} \frac{w_{\sigma_i}}{\sum_{j=i}^{n} w_{\sigma_j}}.$$

Independently and with a completely different interpretation and motivation (related to horse race betting), Plackett (1975) introduced a generalized versions of this model. This is why it is often called the *Plackett-Luce* model in the literature and this is the name we use in the rest of this thesis. In an economics context where elements have features, McFadden (1974a) has shown that it can be seen as a *Multinomial Logistic (MNL)* model and this name is widely used in the economics and Operations Research literature.

The Plackett-Luce model has been extensively studied and under many aspects. It was shown in Block and Marschak (1960) that it is a specific case of the Thurstone model with an implicit proof. Then McFadden (1974a) and Yellott (1977) showed that the Plackett-Luce model is derived from the Thurstone model when the random variables are independent and follow a

Gumbel distribution and reciprocally that the latter is the only distribution that leads to the Plackett-Luce model.

Luce's choice axiom, which can be seen as a probabilistic version of *Independence or Irrelevant Alternatives (IIA)* (one of the axioms in Arrow's impossibility theorem Arrow, 1950), has also been widely discussed (Debreu, 1960; Restle, 1961; Saari, 2005) and several contributions have generalized it (Tversky, 1972; Samuelson, 1985; Gul et al., 2014) or developed statistical methods to overcome its limitations (Ieong et al., 2012; Takahashi and Morimura, 2015).

At last, a wide literature has been devoted to extend the Plackett-Luce model (Henery, 1981; Benter, 1994; Liqun, 2000; Caron and Teh, 2012; Caron et al., 2014) or to design efficient inference procedures (Hunter, 2004; Guiver and Snelson, 2009; Caron and Doucet, 2012; Soufiani et al., 2013b; Maystre and Grossglauser, 2015a).

**The Mallows model.** In Babington Smith (1950), the author proposes to construct a probabilistic model over full rankings as proportional to the product of probabilities over pairwise comparisons. Formally, denoting by $P_{a,b}$ the probability that $a \succ b$ for distinct elements $a, b \in [\![n]\!]$, the model is defined for a full ranking $\sigma \in \mathfrak{S}_n$ by

$$p(\sigma) = C \prod_{1 \leq a < b \leq n} P_{a,b}^{\mathbb{I}\{\sigma(a) < \sigma(b)\}} (1 - P_{a,b})^{\mathbb{I}\{\sigma(a) > \sigma(b)\}}. \tag{2.3}$$

where $C > 0$ is a normalizing constant. Judging this general model too cumbersome, Mallows (1957) proposes several ways to specialize it. The author considers first a Bradley-Terry model for the $P_{a,b}$'s leading to the sometimes called *Mallows-Bradley-Terry model* (we point out that this model for full rankings is different from the Plackett-Luce model). Then he follows a second approach and introduces the *Mallows $\phi$-model*, defined for $\sigma \in \mathfrak{S}_n$ by

$$p(\sigma) = C(\phi) \prod_{1 \leq a < b \leq n} \phi^{\mathrm{sign}(\sigma(b) - \sigma(a))}, \tag{2.4}$$

where $\phi > 0$ and $C(\phi)$ is a normalizing constant. In this model, the ranking $1 \dots n$ is seen as the standard and for any $1 \leq a < b \leq n$, the probability $P_{a,b}$ that the pairwise comparison fits with the standard $a \succ b$ is assumed to be constant, equal to $\phi/(\phi + \phi^{-1})$. More generally when a ranking $\sigma^* \in \mathfrak{S}_n$ is seen as the standard, the Mallows model is usually written as

$$p(\sigma) = C(\gamma) e^{-\gamma d_{KT}(\sigma^*, \sigma)}, \tag{2.5}$$

where $d_{KT}$ is the Kendall's tau distance (see Subsection 2.4.4) and the connection with (2.4) is made with $\gamma = 2\log(\phi)$. Formulation (2.5) shows that the Mallows model is an exponential model and can also be interpreted as a "Gaussian" model where the central permutation $\sigma^*$ corresponds to the mean and the spread parameter $\gamma > 0$ to the inverse of the standard deviation.

Many contributions in the literature have studied or extended the Mallows model (Feigin and Cohen, 1978; Fligner and Verducci, 1986, 1988; Diaconis, 1988; Chung and Marden, 1993; McCullagh, 1993; Lebanon and Lafferty, 2002, 2003; Doignon et al., 2004; Meila and Bao, 2008; Meilă and Bao, 2010; Qin et al., 2010; Plis et al., 2011; Meek and Meila, 2014) or introduced procedures to fit it on the data (Meila et al., 2007; Lu and Boutilier, 2011a; Ceberio et al., 2014). From a social choice perspective, the Mallows model can be interpreted as if there exists an ideal ranking $\sigma^*$ for the social good but voters do not know it and make mistakes in their ballots independently for each pair of candidates with constant probability $1/(1+\phi^2)$ with $\phi$ from (2.4). This interpretation leads Young (1988) to say that the Mallows model was already present as an intuition in Condorcet (1785).

*Remark* 6 (On the Babington Smith model). The general Babington Smith model has been few considered in the literature (Joe and Verducci, 1993). We point out however the approach introduced in Cheng et al. (2012) where the authors study when thresholding probabilities of pairwise comparisons leads to a strict partial order. Though there is no direct relationship with the Babington Smith model, we find the analogy between the two modeling approaches interesting.

**Other models.** We mention some parametric models introduced in the literature independently from the work previously cited. We can identify mainly two families of models. The first one assumes a parametric law for the rank of each element of $[\![n]\!]$ in a full ranking (D'Elia, 2000, 2003; Fasola and Sciandra, 2015). The second family models the generation of a random ranking via an recursive insertion process of the elements of $[\![n]\!]$ into a full ranking (Ailon, 2008; Biernacki and Jacques, 2013).

### 2.5.2 Nonparametric models

In contrast to parametric models, nonparametric models are quite diverse and use various mathematical structures. We try here to describe the main contributions in the literature but we are probably not exhaustive.

**Metrics.** Defining a metric on ranking data enables to formalize a wide variety of problems and perform for instance many statistical tests (Critchlow, 1985; Feigin and Alvo, 1986). The study of such distances is thus already important (Diaconis and Graham, 1977) in particular for partial rankings where their definition can become complex (Critchlow, 1985; Fagin et al., 2006; Webber et al., 2010). Metrics have also been used to construct kernels for estimation (Lebanon and Mao, 2008; Sun et al., 2012).

**Independence modeling.** A natural approach to tackle the complexity of ranking data is to exploit some form of independence assumption. This has been done in the literature using the L-decomposability property defined in Critchlow et al. (1991) (Csiszár, 2008, 2009b), Fourier analysis (Huang et al., 2009b) or the concept of *riffled independence* (Huang and Guestrin, 2009, 2012; Huang et al., 2012).

**Sparsity.** Inspired by the achievements of sparsity-inducing methods in machine learning and compressed sensing, several contributions have introduced inference methods under the assumption of a ranking model with sparse support on $\mathfrak{S}_n$ (Jagabathula and Shah, 2008; Farias et al., 2009; Jagabathula and Shah, 2011; Ding et al., 2015b). One specificity when applied to ranking data is that recovery conditions must cope with the combinatorial structure of $\mathfrak{S}_n$.

**Fourier analysis.** Fourier analysis on rankings exploits the algebraic structure of the symmetric group $\mathfrak{S}_n$. First introduced in the seminal contributions Diaconis (1988, 1989), it has known many developments since then. Using group representation theory, it defines an abstract Fourier transform $\mathcal{F}$ that maps a function $f$ over $\mathfrak{S}_n$ to a collection of Fourier coefficients $\mathcal{F}f = (\widehat{f}(\lambda))_\lambda$ (refer to Subsection 3.2.5 for more details). Though it exhibits some differences with the classic Fourier transform (the two main ones being that the Fourier coefficients $\widehat{f}(\lambda)$ are matrices and that the "frequencies" $\lambda$ are not numbers) it shares some fundamental properties: $\mathcal{F}$ is an isometry and turns convolution product into pointwise product. Many contributions have therefore

used Fourier analysis to design statistical procedures, such as nonlinear approximation (Diaconis, 1989; Lawson et al., 2006), band-limited approximation (Huang et al., 2009a; Irurozki et al., 2011), kernel methods (Kondor and Barbosa, 2010), phase-magnitude decomposition (Kakarala, 2011, 2012) or multiresolution decomposition (Kondor and Dempsey, 2012).

**Markov bases.** Several statistical tests on data require to sample from conditional distributions. Markov bases are an efficient tool for this purpose and several contributions in the literature have provided methods to construct them for different probabilistic models over rankings (Diaconis and Sturmfels, 1998; Diaconis and Eriksson, 2006; Csiszár, 2009a; Sturmfels and Welker, 2012).

**Linear Ordering Polytope.** Let $P_{a,b}$ denote the probability distribution on the pairwise comparison between elements $a$ and $b$ induced by the probability distribution $p$ over $\mathfrak{S}_n$ (see Section 3.1 for the formal definitions, $P_{a,b}$ is called a pairwise marginal of $p$). The admissible region for the vector $(P_{a,b})_{1\leq a<b\leq n} \in [0,1]^{n(n-1)/2}$ is a convex polytope called the *Linear Ordering Polytope*. Many contributions in the literature have studied its geometrical properties and its relationships with ranking models (Reinelt, 1985; Grötschel et al., 1985; Cohen and Falmagne, 1990; Suck, 1992; Fishburn, 1992; Koppen, 1995; Zhang, 2004).

**Nonparametric modeling of pairwise comparisons.** Several contributions have also consider specifically the nonparametric modeling of pairwise comparisons. Among them the HodgeRank framework, introduced in Jiang et al. (2011b) and then further developed (Xu et al., 2012; Dalal et al., 2012; Osting et al., 2013), exploits the topological structure of the pairwise comparison graph. A connection with our work is established in Subsection 6.3.1. Other approaches include for instance approximation of pairwise comparison matrices (Koczkodaj and Orlowski, 1997; Chu, 1998; Koczkodaj and Orłowski, 1999; Dopazo and González-Pachón, 2003; Fülöp, 2008) or probabilistic modeling (Volkovs and Zemel, 2014; Shah et al., 2015a).

# Chapter 3

# Motivations for a new representation

In this chapter we describe in details the motivations for the present work. The first one is the statistical analysis of incomplete rankings. Section 3.1 states the problem formally and highlights the challenges. The other motivation is more general: to localize the parts of information involved in relative marginals. It is explained in Section 3.2.

## Contents



## 3.1 Analysis of incomplete rankings

The practical motivation for this thesis is the application to the analysis of incomplete rankings. This section states the problem and its associated challenges in details.

### 3.1.1 Context

As described in Section 2.2, many modern applications naturally involve ranking data analysis. For instance in a recommendation setting, ranking data represent the users preferences and the

ordered list of recommendations. The ideal situation for such an application would be to know the probability

$$\text{“}Prob(\text{ ranking } \pi \text{ over } A \mid \text{subset of items } A;\ \text{user } u;\ \text{context } c)\text{”}$$

of a ranking $\pi$ that user $u$ would affect to items of $A$ in context $c$. In the footsteps of the ranking data analysis literature, the natural theoretical approach would be to derive these probabilities from ranking models over $\mathfrak{S}_n$ (see Subsection 3.1.2 for the definition). These applications however occur in a large-scale setting: the number $n$ of elements is typically around $10^4$ or $10^6$. They therefore present daunting statistical and computational challenges and this is why contributions in the literature have considered either simpler problems or restricted models (see Section 2.3).

Generic probabilistic modeling should nonetheless enable more flexibility and thus lead to better result, if it is tractable. Fortunately, though $n$ can be very large in these applications, the size of the subset $A$ is usually small, typically around 10. Users express their preferences on small subsets of items and only look at a small number of recommendations. The number of parameters to capture the variability of the data should thus be much more manageable and statistical procedures should be able to capture it.

Such an approach inscribes itself in the statistical analysis of incomplete rankings. Perhaps surprisingly, very few contributions have been devoted to this subject, while the analysis of full rankings, partial rankings or pairwise comparisons have been extensively studied (see Chapter 2). Besides parametric models (see Subsection 3.1.5 for a description of their application), we are only aware of three nonparametric approaches that can handle incomplete rankings, namely those introduced in Yu et al. (2002), Kondor and Barbosa (2010) and Sun et al. (2012) in order to perform statistical tests, estimation and prediction respectively. The principles underlying these approaches are described at length in Subsection 3.1.5.

The purpose of the present work is to introduce a new representation for incomplete rankings that enables to construct flexible probabilistic models and associated statistical procedures (refer to Chapters 4 and 5). In the rest of this section, we properly define the problem and explain the associated challenges.

### 3.1.2 Ranking model and consistency assumption

A *ranking model* is a family of probability distributions that characterize the variability of a statistical population of rankings. In the case of full rankings, the statistical population is only composed of random permutations, and a ranking model reduces to one probability distribution $p$ over the symmetric group $\mathfrak{S}_n$. But when one considers partial or incomplete rankings, they usually are of various types, and the global variability of the statistical population is characterized by a family of probability distributions, one over the rankings of each type. In the case of top-$k$ rankings for instance, the number $k$ usually varies from 1 to $n-1$ between observations, and the global variability of the statistical population is characterized by a family $(P_k)_{1\leq k\leq n-1}$ where for each $k \in \{1,\dots,n-1\}$, $P_k$ is a probability distribution over the set of $k$-tuples with distinct elements (see Busse et al., 2007, for instance).

Incomplete rankings are rankings on subsets of elements. The varying parameter in a statistical population of incomplete rankings is thus the subset of elements involved in each ranking. A ranking model for incomplete rankings is then a family $(P_A)_{A\in\mathcal{P}(\llbracket n\rrbracket)}$ where for each $A \in \mathcal{P}(\llbracket n\rrbracket)$, $P_A$ is a probability distribution over the set $\Gamma(A)$ of rankings on $A$.

*Example* 7. For $n = 3$,

$$\underset{\Gamma(\{1,2\})}{\{12,21\}} \sqcup \underset{\Gamma(\{1,3\})}{\{13,31\}} \sqcup \underset{\Gamma(\{2,3\})}{\{23,32\}} \sqcup \underset{\Gamma(\{1,2,3\})\equiv\mathfrak{S}_3}{\{123,132,213,231,312,321\}} = \Gamma_3$$

so that a ranking model for incomplete rankings on $[\![3]\!]$ is a family $(P_{\{1,2\}}, P_{\{1,3\}}, P_{\{2,3\}}, P_{\{1,2,3\}})$.

If there were no relationship between the different probability distributions of a ranking model, the statistical analysis of partial and/or incomplete rankings would boil down to independent analyses for each type of ranking. Yet one should be able to transfer information from the observation of one type of ranking to another. In a context of top-$k$ rankings analysis, if for instance element $a$ appears frequently in top-1 rankings, it is natural to expect that it is ranked in high position in top-$k$ rankings with larger values of $k$, and reciprocally, if it is usually ranked high in top-$k$ rankings, then its probability of being top-1 should be high. The same intuition holds for incomplete rankings. If element $a$ is usually preferred to element $b$ in pairwise comparisons then rankings on $\{a, b, c\}$ that place $a$ before $b$ should have higher probabilities than the others. Reciprocally if such rankings appear more frequently than the others, then element $a$ should be preferred to element $b$ with high probability in a pairwise comparison.

The literature on ranking data analysis generally makes one fundamental assumption: the observed rankings in a statistical population of interest are induced by full rankings drawn from a single probability distribution $p$ over $\mathfrak{S}_n$. Permutation $\sigma \in \mathfrak{S}_n$ induces ranking $\prec$ or equivalently is a linear extension of ranking $\prec$ if for all $a, b \in [\![n]\!]$, $a \succ b \Rightarrow \sigma(a) < \sigma(b)$. The probability that a random permutation $\Sigma$ drawn from $p$ induces a ranking $\prec$ is thus equal to

$$\mathbb{P}\left[\Sigma \in \mathfrak{S}_n(\prec)\right] = \sum_{\sigma \in \mathfrak{S}_n(\prec)} p(\sigma), \tag{3.1}$$

where $\mathfrak{S}_n(\prec)$ is the set of linear extensions of $\prec$. The *consistency assumption* then stipulates that the probability distributions of a ranking model are all given by Eq. (3.1), thus forming a projective family of distributions. For instance, the set of linear extensions of the top-$k$ ranking $a_1 \succ \cdots \succ a_k \succ$ *the rest*, where $k \in \{1, \ldots, n-1\}$ and $a_1, \ldots, a_k$ are distinct elements in $[\![n]\!]$, is equal to $\{\sigma \in \mathfrak{S}_n \mid \sigma^{-1}(1) = a_1, \ldots, \sigma^{-1}(k) = a_k\}$. The probability $P_k(a_1, \ldots, a_k)$ is thus given by

$$P_k(a_1, \ldots, a_k) = \mathbb{P}\left[\Sigma^{-1}(1) = a_1, \ldots, \Sigma^{-1}(k) = a_k\right] = \sum_{\substack{\sigma \in \mathfrak{S}_n \\ \sigma^{-1}(1)=a_1, \ldots, \sigma^{-1}(k)=a_k}} p(\sigma).$$

A permutation $\sigma$ induces an incomplete ranking $\pi$ on $A \in \mathcal{P}([\![n]\!])$ if it ranks the elements of $A$ in the same order as $\pi$, that is if $\sigma(\pi_1) < \cdots < \sigma(\pi_{|\pi|})$. More generally, we say that word $\pi'$ is a subword of word $\pi$ if there exist indexes $1 \leq i_1 < \cdots < i_{|\pi'|} \leq |\pi|$ such that $\pi' = \pi_{i_1} \ldots \pi_{i_{|\pi'|}}$, and we write $\pi' \subset \pi$. Hence, permutation $\sigma$ induces ranking $\pi$ if and only if $\pi \subset \sigma$. In addition, it is clear that for a word $\pi \in \Gamma_n$ and a subset $A \in \mathcal{P}(c(\pi))$, there exists a unique subword of $\pi$ of content $A$. We denote it by $\pi_{|A}$ and call it the induced ranking of $\pi$ on $A$. The set of linear extensions of a ranking $\pi \in \Gamma(A)$ with $A \in \mathcal{P}([\![n]\!])$ is then $\mathfrak{S}_n(\pi) = \{\sigma \in \mathfrak{S}_n \mid \pi \subset \sigma\} = \{\sigma \in \mathfrak{S}_n \mid \sigma_{|A} = \pi\}$ and the probability $P_A(\pi)$ is given by

$$P_A(\pi) = \mathbb{P}\left[\Sigma(\pi_1) < \cdots < \Sigma(\pi_{|\pi|})\right] = \sum_{\sigma \in \mathfrak{S}_n(\pi)} p(\pi) = \sum_{\substack{\sigma \in \mathfrak{S}_n \\ \pi \subset \sigma}} p(\sigma) = \sum_{\substack{\sigma \in \mathfrak{S}_n \\ \sigma_{|A} = \pi}} p(\sigma). \tag{$*$}$$

*Example* 8. Let $n = 3$. For $\sigma = 231$, one has $\sigma_{|\{1,2\}} = 21$, $\sigma_{|\{1,3\}} = 31$ and $\sigma_{|\{2,3\}} = 23$. For $A = \{1, 3\}$ and $\pi = 31$, one has

$$P_{\{1,3\}}(31) = \mathbb{P}\left[\Sigma(3) < \Sigma(1)\right] = p(231) + p(321) + p(312).$$

We call Eq. ($*$) the *consistency assumption* for the statistical analysis of incomplete rankings. It implies that all the $P_A$'s in the ranking model are *marginal distributions* of the same probability distribution $p$ over $\mathfrak{S}_n$. Abusively, $p$ is also called the ranking model thereafter .

We also extend the definition of a marginal to any function of incomplete rankings. As $\Gamma_n = \bigsqcup_{A \in \mathcal{P}(\llbracket n \rrbracket)} \Gamma(A)$, we embed all the spaces $L(\Gamma(A))$ into $L(\Gamma_n)$, identifying a function $F$ on $\Gamma(A)$ to the function $f$ on $\Gamma_n$ equal to $F$ on $\Gamma(A)$ and to 0 outside of $\Gamma(A)$. One thus has $L(\Gamma_n) = \bigoplus_{A \in \mathcal{P}(\llbracket n \rrbracket)} L(\Gamma(A))$.

**Definition 9** (Marginal operator)**.** The marginal operator on a subset $A \in \mathcal{P}(\llbracket n \rrbracket)$ is the operator $M_A : L(\Gamma_n) \to L(\Gamma(A))$ defined for any $F \in L(\Gamma_n)$ by

$$M_A F(\pi) = \sum_{\sigma \in \Gamma_n,\ \pi \subset \sigma} F(\sigma) \qquad \text{for } \pi \in \Gamma(A). \tag{3.2}$$

Notice that one has in particular $M_A p = P_A$ for all $A \in \mathcal{P}(\llbracket n \rrbracket)$ and $M_A F = 0$ if $F \in L(\Gamma(B))$ with $A \notin \mathcal{P}(B)$.

*Remark* 10 (On the consistency assumption)*.* In this remark we discuss the origin of the consistency assumption in the literature. It first appeared with the study of the relation between choice models and ranking models in Georgescu-Roegen (1958) and Luce (1959) (see Luce, 1977). A choice model is a family $(C_A)_{A \in \mathcal{P}(\llbracket n \rrbracket)}$ where $C_A$ is a probability distribution on $A$ with $C_A(a)$ representing the probability that $a$ is chosen among $A$ for all $a \in A$. The consistency assumption for choice models is: there exists a probability distribution $p$ over $\mathfrak{S}_n$ such that $C_A(a) = \sum_{\sigma \in \mathfrak{S}_n,\ \sigma_{|A}(a)=1} p(\sigma)$. It was shown in Block and Marschak (1960) that it is equivalent to a general Thurstone model (see Subsection 2.5.1 for the definition). Then, because the latter is already very flexible and provides a nice interpretation, the consistency assumption was largely endorsed in the ranking data analysis literature (Marley, 1968; Regenwetter and Marley, 2001; Ailon, 2008). An enlightening example is the Mallows model (see Subsection 2.5.1 for the definition). It was initially introduced following the approach of Babington Smith (1950) to define a probability distribution $p$ over $\mathfrak{S}_n$ from probabilities on pairwise comparisons. It happens however that this approach does not satisfy the consistency assumption: the initial probabilities on pairwise comparisons used to construct $p$ are not equal to its pairwise marginals (it is easy to check that this can never be the case except for a Dirac distribution on $\mathfrak{S}_n$). Most of the contributions using the Mallows model have then chosen to take the pairwise probabilities equal to the marginals and not the initial ones (see for instance Lu and Boutilier, 2011a). At last, we point out that alternative assumptions have also been considered for pairwise comparisons (see for instance Luce and Suppes, 1965; Fishburn, 1973; Rajkumar and Agarwal, 2014).

### 3.1.3 Probabilistic setting

A dataset of full rankings is naturally modeled as a collection of random permutations $(\Sigma_1, \dots, \Sigma_N)$ drawn IID from a ranking model $p$. The latter thus fully characterizes the statistical population as well as its observation process. This property does not hold true in the statistical analysis of incomplete rankings, where the ranking model characterizes the statistical population, but it does not entirely characterize the generating process of this population. More specifically, it characterizes the variability of the observations on each subset of elements $A \in \mathcal{P}(\llbracket n \rrbracket)$, but it does not account for the variability of the observed subsets of elements.

*Example* 11*.* A ranking model $p$ for incomplete rankings on $\llbracket 3 \rrbracket$ induces the probability distributions $P_{\{1,2\}}$, $P_{\{1,3\}}$, $P_{\{2,3\}}$ and $P_{\{1,2,3\}} = p$. For each $A \in \mathcal{P}(\llbracket 3 \rrbracket)$, a random ranking on $A$ can thus be drawn from the probability distribution $P_A$. But the $P_A$'s do not induce a probability distribution on $\mathcal{P}(\llbracket 3 \rrbracket)$ that would generate the samplings of the subsets $A$.

To model this double variability, we represent the observation of an incomplete ranking by a couple of random variables $(\mathbf{A}, \Pi)$, where $\mathbf{A} \in \mathcal{P}(\llbracket n \rrbracket)$ is the observed subset of elements and

$\Pi \in \Gamma(A)$ is the observed ranking *per se* on this subset of elements. Let $\nu$ be the law of $\mathbf{A}$ over $\mathcal{P}(\llbracket n \rrbracket)$. A dataset of incomplete rankings is then a collection $((\mathbf{A}_1, \Pi^{(1)}), \ldots, (\mathbf{A}_N, \Pi^{(N)}))$ of IID samples of $(\mathbf{A}, \Pi)$ drawn from the following process:

$$\mathbf{A} \sim \nu \qquad \text{then} \qquad \Pi | (\mathbf{A} = A) \sim P_A. \tag{3.3}$$

The interpretation of probabilistic setting (3.3) is that first the subset of elements $\mathbf{A} \in \mathcal{P}(\llbracket n \rrbracket)$ is drawn from $\nu$ and then the ranking $\Pi \in \Gamma(A)$ is drawn from $P_A$. It can be reformulated by exploiting the consistency assumption $(*)$. The latter stipulates that for $A \in \mathcal{P}(\llbracket n \rrbracket)$, the distribution of the random variable $\Pi$ on $\Gamma(A)$ is the same as that of the induced ranking $\Sigma_{|A}$ of a random permutation $\Sigma$ drawn from $p$. A drawing of $(\mathbf{A}, \Pi)$ can thus be reformulated as

$$\Sigma \sim p \qquad \text{then} \qquad \mathbf{A} \sim \nu \qquad \text{and} \qquad \Pi = \Sigma_{|\mathbf{A}}. \tag{3.4}$$

Reformulation (3.4) leads to the following interpretation: first a random permutation $\Sigma \in \mathfrak{S}_n$ is drawn from $p$ then the subset of elements $\mathbf{A}$ is drawn from $\nu$ and the ranking $\Pi$ is set equal to $\Sigma_{|\mathbf{A}}$. The permutation $\Sigma$ can then be seen as a latent variable that expresses the full preference of a user in the statistical population but its observation is censored by $\mathbf{A}$. We point out that this interpretation motivates the broader probabilistic setting introduced in Sun et al. (2012). The authors model more generally the observation of any partial and/or incomplete ranking as the drawing of a latent random permutation $\Sigma$ from $p$ followed by a censoring process that *can depend* on $\Sigma$. In the context of incomplete rankings observation, their probabilistic setting can be defined as: $\Sigma \sim p$ then $\mathbf{A} \sim \nu_\Sigma$ and $\Pi = \Sigma_{|\mathbf{A}}$, where $\nu_\sigma$ is a probability distribution over $\mathcal{P}(\llbracket n \rrbracket)$ for each $\sigma \in \mathfrak{S}_n$. Probabilistic setting (3.3) fits into this broader one by setting all distributions $\nu_\sigma$ equal to $\nu$ or, equivalently, assuming that $\Sigma$ and $\mathbf{A}$ are *independent*.

The independence between $\Sigma$ and $\mathbf{A}$ corresponds to the *missing at random assumption* in the general context of learning from incomplete data (see Ghahramani and Jordan, 1995). This assumption is not realistic in all situations, particularly in settings where the users choose the items on which they express their preferences, their choices being naturally biased by their tastes (see Marlin et al., 2007, for instance). It remains however realistic in many situations where the subset of items proposed to the user is determined by the context: the available items in stock in a specific store or the possible recommendations in a specific area for instance. This assumption is thus made in many contributions of the literature (see for instance Lu and Boutilier, 2014; Rajkumar and Agarwal, 2014; Ding et al., 2015b).

*Remark* 12. We maintain furthermore that making a dependence assumption is incompatible with the principle of the statistical analysis of incomplete rankings. Indeed, the purpose of assuming that $\mathbf{A}$ and the latent variable $\Sigma$ are not independent is to infer from the observation of $\Pi = \Sigma_{|\mathbf{A}}$ some more information on $\Sigma$ than just $\Sigma_{|\mathbf{A}}$. For instance, to model the fact that the expression of user's preferences could be biased by their tastes, one can assume that the full ranking $\Sigma$ is censored to elements that have a low expected rank (meaning that they have a high probability to be ranked in the first positions). The subset of elements $\mathbf{A}$ could then be obtained by sampling elements without replacement from a distribution over $\llbracket n \rrbracket$ of the form $\eta_\sigma(i) \propto e^{-\alpha\sigma(i)}$, where $\alpha \in \mathbb{R}$ is a spread parameter, conditioned upon $\Sigma = \sigma$. The observed ranking $\Pi$ on a subset $A = \{a_1, \ldots, a_k\} \in \mathcal{P}(\llbracket n \rrbracket)$ would then not only provide information on the relative ordering $\Sigma_{|A}$ of the elements of $A$ but even more on their absolute ranks $(\Sigma(a_1), \ldots, \Sigma(a_k))$ in the latent full ranking $\Sigma$. Exploiting this additional information requires to analyze $\Pi$ as a partial ranking. Thus it cannot be done in a setting of statistical analysis of incomplete rankings.

### 3.1.4 Challenges of the statistical analysis of incomplete rankings

We now formalize the general setting for the statistical analysis of incomplete rankings. One observes a dataset $\mathcal{D}_N = ((\mathbf{A}_1, \Pi^{(1)}), \dots, (\mathbf{A}_N, \Pi^{(N)}))$ of $N$ incomplete rankings drawn IID from the process (3.3) with $p$ a ranking model and $\nu$ a probability distribution over $\mathcal{P}(\llbracket n \rrbracket)$. The ranking model $p$ is unknown and the goal is to summarize or recover some part of it. The probability distribution $\nu$ is assumed to be known, it is indeed not the purpose of ranking data analysis in general to infer it (though it may be an interesting problem). It remains however the censoring process that generates the design of observations and we thus call it the *observation design* (by analogy with random design regression in classic statistics). It has a major impact on the parts of $p$ that can be inferred from the dataset $\mathcal{D}_N$ (a deeper analysis is provided in Subsection 3.1.6).

Characterizing separately the variability of the observed subset $\mathbf{A}$ leads to an unexpected analogy with supervised learning: in the couple $(\mathbf{A}, \Pi)$, the subset $\mathbf{A}$ can be seen as an input generated by the distribution $\nu$ and the ranking $\Pi$ can be seen as the output generated by the ranking model $p$ given the input $\mathbf{A}$. Analyzing incomplete rankings data thus requires to face two classical issues in statistical learning, which can be easily formulated in the context of binary classification, the flagship problem in machine-learning theory.

- **Consolidate knowledge on already seen subsets of elements.** For an observed subset $A \in \mathcal{P}(\llbracket n \rrbracket)$, one must consolidate all the observations on $A$ in order to recover a maximum amount of information about $P_A$. The corresponding task in binary classification is to consolidate all the outputs $y$ for a given input $x$ (or very close inputs) that was observed many times, where $x$ and $y$ are the values taken by IID samples of a random couple $(X, Y)$. Its difficulty depends on how much the value $\mathbb{P}[Y = 1 | X = x]$ is close to $1/2$: the closer the more difficult. Analogously, the difficulty of consolidating observations on a given subset of elements $A$ depends on the complexity of the marginal $P_A$. If $P_A$ is a Dirac function, it is easy to recover. If $P_A$ is more complex, its recovery is more challenging.

- **Transfer knowledge to unseen subsets of of elements.** For a new unseen subset, one needs to transfer a maximum amount of acquired information from the observed subsets. In binary classification, one faces an analogous problem when trying to predict the output $y$ related to an input value $x$ never observed before and potentially far from all previously observed inputs. The difficulty of this task then depends on the "regularity" of the function $\eta : x \mapsto \mathbb{P}[Y = 1 | X = x]$: it is easier to infer the value of $\mathbb{P}[Y = 1 | X = x]$ for an unobserved $x$ when $\eta$ is "regular", in the sense that $\eta(x)$ does not vary unexpectedly when $x$ varies. Similarly for incomplete rankings, it is easier to transfer information to an unobserved subset of elements $A$ when the function $B \mapsto P_B$ does not vary unexpectedly when $B$ varies in $\mathcal{P}(\llbracket n \rrbracket)$.

These two tasks require to cope with two different sources of variability and can be tackled independently in a theoretical setting. But in a statistical setting, they must be handled simultaneously in order to best reduce the sampling noise of a dataset $\mathcal{D}_N$. It is better indeed to transfer between subsets information that has been consolidated on each subset and conversely, it is better to consolidate information on a subset with information transferred from other subsets. A major difficulty however remains: incomplete rankings are heterogeneous. They can have different sizes and for a given size they can be observed on different subsets of elements. Consolidating and transferring information for incomplete rankings is thus far from being obvious, and represents the main statistical challenge of the analysis of incomplete rankings.

*Example* 13. Let $n = 4$ and assume that one observes rankings on $\{1, 3\}$, $\{1, 3, 4\}$ and $\{2, 4\}$. Information could be consolidated on each of these three subsets independently and then transferred

to unobserved subsets. It would certainly be more efficient however to consolidate information on these subsets simultaneously, transferring at the same time information between them. The question is now to find a way to achieve this.

The consistency assumption (∗) defines the base structure to transfer information between subsets of elements. Namely for two subsets $A, B \in \mathcal{P}(\llbracket n \rrbracket)$ with $B \subset A$, it stipulates that $M_B P_A = P_B$. The knowledge of $P_A$ thus implies the knowledge of $P_B$. Information must therefore be transferred from $A$ to $B$ through the marginal operator $M_B$. The condition is slightly more subtle in the other direction: information must be transferred from $B$ to $A$ through the constraint on $P_A$ to satisfy $M_B P_A = P_B$. Hence, the knowledge of $P_B$ does not imply the knowledge of $P_A$, but it provides some part of it. More generally, the knowledge of any marginal $P_A$ provides some information on $p$ through the constraint $M_A p = P_A$. How to transfer information from $A$ to a subset $C$ such that neither $C \subset A$ nor $A \subset C$ is however a priori unclear.

*Example* 14. Coming back to the previous example, information on $\{1, 3, 4\}$ should be used to consolidate information on $\{1, 3\}$ through the relationship $M_{\{1,3\}} P_{\{1,3,4\}} = P_{\{1,3\}}$. Information on $\{1, 3\}$ should be used to enforce a constraint in consolidating information on $\{1, 3, 4\}$ through the same relationship. Information on each subset can be used to enforce a constraint on the global ranking model $p$. It is however unclear if or how information should be transferred between $\{2, 4\}$ and $\{1, 3\}$ or $\{1, 3, 4\}$.

In addition to this major statistical challenge, practical applications also raise a great computational challenge. The analysis of incomplete rankings always involve at some point the computation of a marginal of a ranking model. Performed naively using Definition 9, the computation of $M_A p(\pi)$ for $A \in \mathcal{P}(\llbracket n \rrbracket)$ and $\pi \in \Gamma(A)$ requires $n!/|A|!$ operations (the number of full rankings that extend $\pi$). This is by far intractable in practical applications where $|A|$ is around 10 and $n$ is around $10^4$.

### 3.1.5 Limits of existing approaches

We now review the existing approaches in the literature for the statistical analysis of incomplete rankings and outline their limits.

**Parametric models.** The most widely used approaches rely on parametric modeling (see Subsection 2.5.1). One considers a family of models $\{p_\theta \mid \theta \in \Theta\}$, where $\Theta$ is a parameter space, and assumes that $p = p_{\theta^*}$ for a certain $\theta^* \in \Theta$. The goal is then to recover $\theta^*$ from the dataset $\mathcal{D}_N$. One standard method is to take the parameter that maximizes the likelihood of the model on the dataset. For $\theta \in \Theta$, let $\mathbb{P}_\theta$ be the distribution of a random permutation $\Sigma$ corresponding to the ranking model $p_\theta$, that is to say the distribution defined by $\mathbb{P}_\theta[\Sigma \in S] = \sum_{\sigma \in S} p_\theta(\sigma)$ for any subset $S \subset \mathfrak{S}_n$. The relevance of a candidate model $p_\theta$ on the dataset $\mathcal{D}_N$ is thus measured through the conditional likelihood

$$\mathcal{L}(\theta | \mathbf{A}_1, \dots, \mathbf{A}_N) = \prod_{i=1}^{N} \mathbb{P}_\theta \left[ \Sigma_{|\mathbf{A}_i} = \Pi^{(i)} \right] = \prod_{i=1}^{N} M_{\mathbf{A}_i} p_\theta \left( \Pi^{(i)} \right).$$

One then compute $\widehat{\theta}_N = \operatorname{argmax}_{\theta \in \Theta} \mathcal{L}(\theta | \mathbf{A}_1, \dots, \mathbf{A}_N)$ exactly or approximately and uses the ranking model $\widehat{p}_N := p_{\widehat{\theta}_N}$. In this approach, the consolidation of information is performed implicitly through the selection of the ranking model from the family $\{p_\theta \mid \theta \in \Theta\}$ that best explains the data. It is then transferred to any subset of elements $B \in \mathcal{P}(\llbracket n \rrbracket)$ through the marginal $M_B \widehat{p}_N$. The computational challenge is easily overcome when using the Plackett-Luce model because the marginals of the latter have a closed-form expression. The Thurstone model

is naturally fitted by breaking the incomplete rankings into pairwise comparisons. It is much less straightforward for the Mallows model, but a dedicated method was introduced in Lu and Boutilier (2011a). From a global point of view, approaches based on a parametric model have the advantage to offer a simple framework for all applications of the statistical analysis of incomplete rankings. Their major drawback however is to rely on a rigid assumption on the form of the ranking model, which is rarely satisfied in practice.

**Nonparametric methods based on identifying an incomplete ranking with the set of its linear extensions.** The three nonparametric methods introduced in the literature to analyze incomplete rankings all face the heterogeneity of incomplete rankings the same way: they represent an incomplete ranking $\pi \in \Gamma_n$ by the set of its linear extensions $\mathfrak{S}_n(\pi) \subset \mathfrak{S}_n$. Yu et al. (2002) generalize a distance $d$ on $\mathfrak{S}_n$ to a distance $d^*$ on $\Gamma_n$ by setting $d^*(\pi, \pi')$ proportional to $\sum_{\sigma \in \mathfrak{S}_n(\pi)} \sum_{\sigma' \in \mathfrak{S}_n(\pi')} d(\sigma, \sigma')$ for two incomplete rankings $\pi, \pi' \in \Gamma_n$ and use it to perform statistical tests. In Sun et al. (2012), the Kendall's tau distance is generalized in the same way and then used to define a kernel-based estimator of $p$. Finally, Kondor and Barbosa (2010) define kernels on $\Gamma_n$ based, for two incomplete rankings $\pi, \pi' \in \Gamma_n$, on the Fourier transform of the indicator functions of the sets $\mathfrak{S}_n(\pi)$ and $\mathfrak{S}_n(\pi')$. Broadly speaking, these three approaches transfer information between different incomplete rankings through a given similarity measure *between their sets of linear extensions*. They overcome some part of the computational challenge through explicit simplifications of the extended distance $d^*$ or the Fourier transform of the indicator function of an incomplete ranking. They are however fundamentally biased. To best illustrate this point, let us consider the following estimator:

$$\widehat{p}_N = \frac{1}{N} \sum_{i=1}^{N} \frac{|\mathbf{A}_i|!}{n!} \mathbb{1}_{\mathfrak{S}_n(\Pi^{(i)})}. \tag{3.5}$$

It corresponds to the natural empirical estimator of $p$ when one represents an incomplete ranking by the set of its linear extensions. In this representation indeed, one considers that the observation of an incomplete ranking $\Pi$ indicates that the underlying permutation $\Sigma$ should belong to $\mathfrak{S}_n(\Pi)$. The amount of knowledge about $\Sigma$ is thus modeled by the uniform distribution on $\mathfrak{S}_n(\Pi)$. The estimator $\widehat{p}_N$ is then the average of the uniform distributions over the sets $\mathfrak{S}_n(\Pi^{(i)})$ for $i \in \{1, \dots, N\}$. As stated in the following proposition, it is always strongly biased, except in a few specific situations, irrelevant in practice.

**Proposition 15.** *Let $N \geq 1$ and $\widehat{p}_N$ be the estimator defined by equation (3.5). Then for any $\sigma \in \mathfrak{S}_n$,*

$$\mathbb{E}\left[\widehat{p}_N(\sigma)\right] = \sum_{\sigma' \in \mathfrak{S}_n} \left( \sum_{A \in \mathcal{P}(\llbracket n \rrbracket)} \nu(A) \frac{|A|!}{n!} \mathbb{I}\{\sigma'_{|A} = \sigma_{|A}\} \right) p(\sigma').$$

*Proof.* Using the reformulation (3.4) of the data generating process producing the observations, one has for any $\sigma \in \mathfrak{S}_n$

$$\mathbb{E}\left[\widehat{p}_N(\sigma)\right] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[ \frac{|\mathbf{A}_i|!}{n!} \mathbb{I}\{\Sigma_{|\mathbf{A}_i} = \sigma_{|\mathbf{A}_i}\} \right] = \sum_{A \in \mathcal{P}(\llbracket n \rrbracket)} \nu(A) \frac{|A|!}{n!} \sum_{\sigma' \in \mathfrak{S}_n} p(\sigma') \mathbb{I}\{\sigma'_{|A} = \sigma_{|A}\}.$$

A simple sum inversion concludes the proof. □

Proposition 15 says that unless $p$ is a Dirac distribution (which is a too restrictive assumption) or $\nu$ is solely concentrated on $\llbracket n \rrbracket$ (which boils down to statistical analysis on full rankings), $\mathbb{E}\left[\widehat{p}_N(\sigma)\right]$ is fundamentally different from $p(\sigma)$ for $\sigma \in \mathfrak{S}_n$.

*Example* 16. Let $n = 4$ and $\nu$ with support $\{\{1,3\},\{2,4\},\{1,3,4\}\}$. Then for any $N \geq 1$,

$$\begin{aligned}\mathbb{E}\left[\widehat{p}_N(2134)\right] = &\frac{\nu(\{1,3\})}{12}\Big[p(2413)+p(4213)+p(2134)+p(4132)+p(1324)+p(1342)\Big]\\ &+\frac{\nu(\{2,4\})}{12}\Big[p(1324)+p(3124)+p(1243)+p(3241)+p(2413)+p(2431)\Big]\\ &+\frac{\nu(\{1,3,4\})}{4}\Big[p(2134)+p(1342)\Big].\end{aligned}$$

We point out that Proposition 15 says more specifically that $\widehat{p}_N$ is actually an unbiased estimator of $T_\nu p$, where $T_\nu$ is the matrix of similarity defined by

$$T_\nu(\sigma,\sigma') = \sum_{A\in\mathcal{P}(\llbracket n\rrbracket)} \nu(A)\frac{|A|!}{n!}\mathbb{I}\{\sigma_{|A} = \sigma'_{|A}\} \qquad \text{for } \sigma,\sigma' \in \mathfrak{S}_n,$$

In particular, if $\nu$ is the uniform distribution over the pairs of $\llbracket n\rrbracket$, $T_\nu(\sigma,\sigma')$ simply reduces to an affine transform of the Kendall's tau distance between $\sigma$ and $\sigma'$.

**Learning from incomplete rankings as a regularized inverse problem.** A general framework for the statistical analysis of incomplete rankings could take the paradigmatic form of a regularized inverse problem. Assume first that one knows exactly some of the marginals of the ranking model $p$, for a collection of subsets $\mathcal{A} \subset \mathcal{P}(\llbracket n\rrbracket)$. He could try to recover $p$ through the minimization problem

$$\min_{\substack{q:\mathfrak{S}_n\to\mathbb{R}\\ q\geq 0\\ \sum_{\sigma\in\mathfrak{S}_n} q(\sigma)=1}} \Omega(q) \qquad \text{subject to} \qquad M_A q = P_A \text{ for all } A\in\mathcal{A}, \tag{3.6}$$

where $\Omega$ is a penalty function that measures a certain level of regularity, so that $p$ should be a solution of (3.6). Information from the $P_A$'s would then be transferred to an unknown subset $B \in \mathcal{P}(\llbracket n\rrbracket)$ through the computation of $M_B p^*$, where $p^*$ is an exact or approximate solution of (3.6). In a statistical setting, one cannot know exactly the marginals of $p$. The natural extension is then to consider the naive empirical estimator defined for an observed subset $A$ by

$$\widehat{P_A}(\pi) = \frac{|\{1\leq i\leq N \mid \Pi^{(i)} = \pi\}|}{\widehat{N}_A} \qquad \text{for } \pi\in\Gamma(A), \tag{3.7}$$

where $\widehat{N}_A$ denotes the number of times that $A$ was observed in $\mathcal{D}_N$, and to consider the following generic minimization problem

$$\min_{\substack{q:\mathfrak{S}_n\to\mathbb{R}\\ q\geq 0\\ \sum_{\sigma\in\mathfrak{S}_n} q(\sigma)=1}} \sum_{\substack{A\in\mathcal{P}(\llbracket n\rrbracket)\\ \widehat{N}_A>0}} \frac{\widehat{N}_A}{N}\Delta_A\left(M_A q, \widehat{P_A}\right) \quad + \quad \lambda_N\Omega(q), \tag{3.8}$$

where $\Delta_A$ is a dissimilarity measure between two probability distributions over $\Gamma(A)$[1] and $\lambda_N$ is a regularization parameter. Information is then simultaneously consolidated on the observed subsets into an exact or approximate solution $\widehat{p}_N$ and can then be transferred to unobserved subsets by computing $M_B\widehat{p}_N$.

Though this approach is quite common in the machine learning literature, where $\Omega(q)$ typically enforces the sparsity of $q$ in a certain basis, it has been applied to the ranking literature

[1] One can take for instance an $l^p$ norm or the Kullback-Leibler divergence.

only in a few contributions. In Jagabathula and Shah (2011) for instance, the problem of recovering the ranking model $p$ from the observation of its first-order absolute marginals $\mathbb{P}[\Sigma(i) = j]$ for $i \in [\![n]\!]$ and $j \in \{1, \dots, n\}$ (see Section 3.2 for the general definition) is considered under a sparsity assumption over $\mathfrak{S}_n$. A maximal entropy assumption is made in Ammar and Shah (2012) in order to recover the ranking model $p$ either from its first-order absolute marginals or from its pairwise relative marginals $\mathbb{P}[a \succ b]$ for $a, b \in [\![n]\!]$, $a \neq b$.

In the setting of the statistical analysis of incomplete rankings, this approach has the advantage to allow less restrictive assumptions than parametric modeling and to avoid the bias of the aforementioned nonparametric approaches. It suffers however from a major drawback: it requires to compute the marginal operators. It is therefore inapplicable on practical datasets if this computation is performed naively through Definition 9.

All the existing approaches follow the same two steps: first, information from the dataset $\mathcal{D}_N$ is consolidated and transferred into a ranking model $\widehat{p}_N \in L(\mathfrak{S}_n)$. Then it can be transferred to any subset of elements $B \in \mathcal{P}([\![n]\!])$ through the marginal $M_B \widehat{p}_N$. This is of course the most natural method to exploit the consistency assumption $(*)$ and try to overcome the statistical challenge of the analysis of incomplete rankings. It does not provide however any help to overcome the challenge of the computation of the marginal. This is why each approach requires a specific trick to be applicable.

In Chapter 5 we introduce a general framework that enables to handle both the computational and statistical challenges of the analysis of incomplete rankings. Instead of consolidating information into a ranking model $\widehat{p}_N \in L(\mathfrak{S}_n)$, observations are first represented into a feature space, which we call the MRA representation, fitted to exploit the consistency assumption and to compute the marginal operator efficiently. The framework then provides many possibilities to consolidate and transfer information in this feature space. The MRA representation is entirely model-free, it simply arises from the natural multiscale structure of the marginal operators and its algebraic and topological properties (see Chapter 4).

### 3.1.6 Impact of the observation design

Depending on the application, the observation design $\nu$ may or may not be known. In any case, as explained in Subsection 3.1.4, it is not the goal of the statistical analysis of incomplete rankings to learn it. It is rather seen as a parameter that adds some noise to the observations through the censoring process (3.4). It has nonetheless a direct impact on the complexity of the analysis, both on the statistical and computational points of view, especially through its support $\mathcal{A} = \{A \in \mathcal{P}([\![n]\!]) \mid \nu(A) > 0\}$. The first one is on the number of parameters required to store a dataset $\mathcal{D}_N$.

**Lemma 17.** *The number of parameters required to store the dataset $\mathcal{D}_N$ is upper bounded by* $\min(N, \sum_{A \in \mathcal{A}} |A|!)$.

Refer to the Appendix for the proof of Lemma 17. The number $\min(N, \sum_{A \in \mathcal{A}} |A|!)$ given by Lemma 17 is a measure of the "complexity" of the dataset $\mathcal{D}_N$, in the sense that any procedure that exploits all the information contained in $\mathcal{D}_N$ will necessarily require at least as many operations. Notice that the number $\sum_{A \in \mathcal{A}} |A|!$ is entirely characterized by $\mathcal{A}$. It increases both with its "spread" $|\mathcal{A}|$ and its "depth" $K = \max_{A \in \mathcal{A}} |A|$, and is bounded by $|\mathcal{A}| \times K!$. In particular if $\mathcal{A} = \{A \in \mathcal{P}([\![n]\!]) \mid |A| \leq K\}$ then this bound is of order $O(K!\, n^K)$. Figures 3.1 and 3.2 show two examples of $n = 5$ with the associated number $\sum_{A \in \mathcal{A}} |A|!$. The elements in $\mathcal{A}$ are in black whereas the elements of $\mathcal{P}([\![n]\!]) \setminus \mathcal{A}$ are in gray.

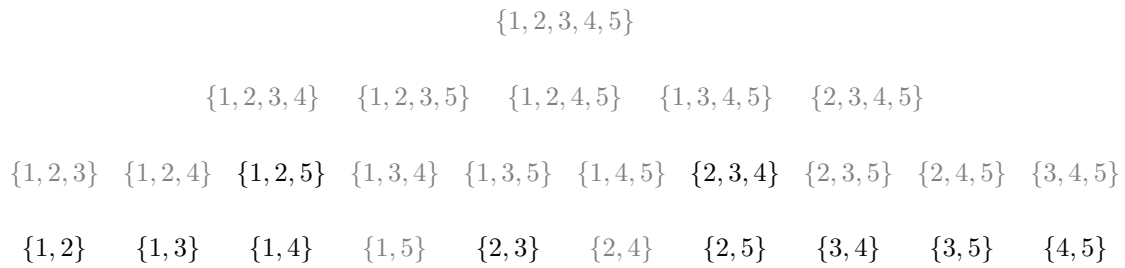



Figure 3.1: Example of an observation design $\mathcal{A}$ for $n = 5$, $\sum_{A\in\mathcal{A}} |A|! = 28$

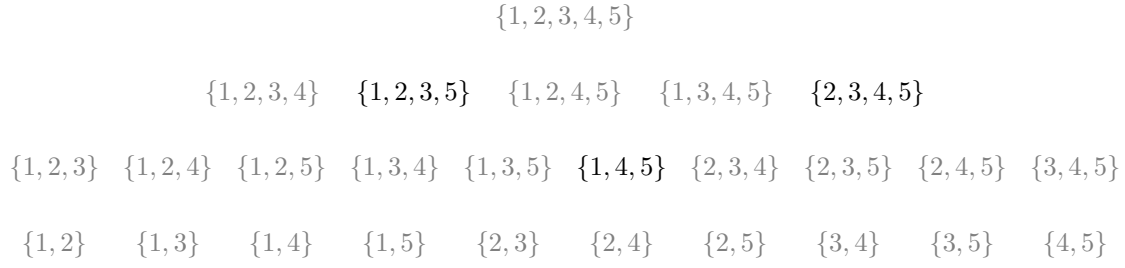



Figure 3.2: Example of an observation design $\mathcal{A}$ for $n = 5$, $\sum_{A\in\mathcal{A}} |A|! = 54$

The observation design also impacts the accessible amount of information about the ranking model $p$. Indeed, if one makes a structural assumption on $p$ and seeks to recover some part of it from the observation of incomplete rankings drawn from (3.3), the complexity of this task will significantly depend on the interplay between $p$ and $\nu$.

*Example* 18. As a toy example, consider the very simple case where one observes the exact induced rankings of one full ranking $\pi^*$ on $[\![5]\!]$, on the subsets $\{1, 2, 3\}$, $\{3, 4\}$ and $\{4, 5\}$. The goal is then to recover the ranking model $p = \delta_{\pi^*}$ through the observation design $\mathcal{A} = \{\{1, 2, 3\}, \{3, 4\}, \{4, 5\}\}$. If $\pi^* = 12345$, then the observed induced rankings are 123, 34 and 45. It happens that there is only one full ranking on $[\![5]\!]$ that induces these three rankings, namely 12345, and $\pi^*$ is recovered with certainty. Now, if $\pi^* = 24153$ for instance, the observed rankings are 213, 43 and 45. In that case, there are twelve full rankings on $[\![5]\!]$ that can induce these three rankings. The amount of information provided by these observations is therefore not sufficient for recovering $\pi^*$.

In a general context, one may assume that $p$ has a more general structure than a Dirac function on $\Gamma([\![n]\!])$ and that the observations are made in the presence of a statistical noise. But the principle illustrated by Example 18 remains valid. Quantifying the amount of accessible information with respect to the interplay between $p$ and $\mathcal{A}$ is however not obvious because the latter is of a complex combinatorial nature. When no structural assumption is made on $p$, the accessible information can be characterized exactly through the MRA representation, by Theorem 99 in Section 5.1.

## 3.2 Localization of relative rank information

The other motivation for this thesis is more theoretical. It is about providing the tools to exploit the multiscale structure of marginals and "localize relative rank information". This section explains the related concepts in details and shows the interest for such a contribution.

### 3.2.1 Marginals of a ranking model

In Section 3.1, we showed that under the natural consistency assumption ($*$) and probabilistic setting (3.3), the marginals of the ranking model $p$ are the only statistics one can access when observing incomplete rankings. This would also be true when observing top-$k$ or more general partial rankings, once the analogues for marginals, consistency assumption and probabilistic setting are defined. In the other way round, even if one knows the ranking model $p$, marginals provide useful summary statistics for it. This interpretation is developed at length for marginals associated to partial rankings in Diaconis (1989) or Huang (2011) for instance.

In both cases, a natural question is: how much information does a marginal of $p$, or more generally a collection of marginals of $p$, contain about $p$? To answer this question in general, one should first define what is a marginal of $p$. This requires however the introduction of many concepts and notations, and it is not necessary for the purpose of this thesis. Instead, we give here an informal definition for general marginals and a proper definition for marginals associated to partial rankings, that we call *absolute marginals*, in Subsection 3.2.2.

Permutations, or full rankings, are complex objects. There are indeed $n!$ different full rankings of $[\![n]\!]$ so the specification of one is the specification of one element among $n!$ possibilities. The total part of information related to this knowledge can however be decomposed in simpler parts, each accessible by a "query". For instance one can ask about a full ranking $\sigma \in \mathfrak{S}_n$: what is the rank $\sigma(a)$ of element $a$? What are the two first elements $\{\sigma_1, \sigma_2\}$? What is the relative order $\sigma_{|\{a,b\}}$ of elements $a$ and $b$?

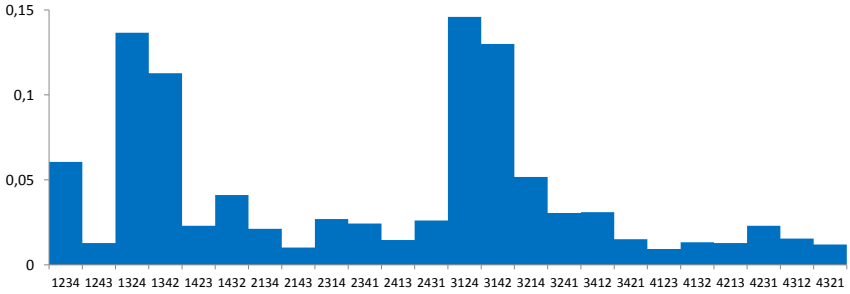
*Example* 19 (Queries on a full ranking). Here are the answers to some queries about the full ranking

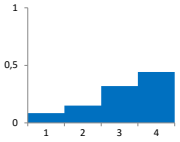
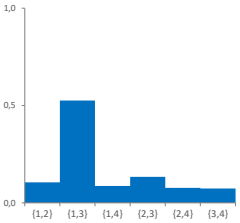
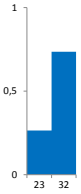
$$\sigma = 2143$$

| Query | Answer |
|---|---|
| What is the rank $\sigma(4)$ of element 4? | $\mathbf{3}$ |
| What are the two first elements $\{\sigma_1, \sigma_2\}$? | $\{\mathbf{2}, \mathbf{1}\}$ |
| What is the relative order $\sigma_{\|\{2,3\}}$ of elements 2 and 3? | $\mathbf{2} \prec \mathbf{3}$ |

In the case of a random full ranking $\Sigma$ drawn from a probability distribution $p$ over $\mathfrak{S}_n$, the analogous queries are: what is the law of the rank $\Sigma(a)$ of element $a$? What is the law of the two first elements $\{\Sigma_1, \Sigma_2\}$? What is the law of the relative order $\Sigma_{|\{a,b\}}$ of elements $a$ and $b$?

*Example* 20 (Queries on a random full ranking). Here are the answers to the queries analogous to the ones of Example 19, for a random full ranking $\Sigma$ drawn from the distribution below ($p$ from the German dataset).




| Query | Answer |
|---|---|
| What is the law of $\Sigma(4)$? |  |
| What is the law of $\{\Sigma_1, \Sigma_2\}$? |  |
| What is the law of $\Sigma_{\|\{2,3\}}$? |  |







The answers to queries on $\Sigma$ are the marginals of $p$. They are probability distributions obtained from $p$. They include the marginals on subsets defined in Section 3.1. Indeed for $A = \{a_1, \ldots, a_k\} \in \mathcal{P}([\![n]\!])$, the answer to the query "what is the law of the relative order $\Sigma_{|A}$ of elements $a_1, \ldots, a_k$?" is by definition $M_A p$. This informal definition of general marginals also includes the class usually considered in the literature, that we call *absolute marginals*.

### 3.2.2 Absolute marginals

What we call *absolute marginals* correspond to the marginals usually considered in the literature (see for instance Diaconis, 1988). We give them a specific name here to differentiate from the marginal operators $M_A$, and explain it in Subsection 3.2.3. Absolute marginals are related to queries about partitions of $[\![n]\!]$.

**Definition 21** (Partition of $[\![n]\!]$)**.** A partition of $[\![n]\!]$ is a collection $\mathcal{B} = \{B_1, \ldots, B_r\}$ of non-empty and two-by-two disjoint subsets of $[\![n]\!]$ such that $\bigsqcup_{i=1}^{r} B_i = [\![n]\!]$, with $1 \leq r \leq n$. We denote by $\mathrm{Part}([\![n]\!])$ the set of all partitions of $[\![n]\!]$.

For a partition $\mathcal{B} = \{B_1, \ldots, B_r\}$ of $[\![n]\!]$, one can ask the following queries about a fixed

permutation $\sigma \in \mathfrak{S}_n$ (respectively a random permutation $\Sigma$ drawn from the ranking model $p$):

$$\begin{aligned}&\text{What is the set of ranks } \sigma(B_i) \text{ (respectively the law of the set of ranks } \Sigma(B_i))\\ &\text{of the elements of each subset } B_i?\end{aligned} \tag{3.9}$$

$$\begin{aligned}&\text{What is the set of elements } \sigma^{-1}(B_i) \text{ (respectively the law of the set of elements } \Sigma^{-1}(B_i))\\ &\text{ranked at positions } b_{i,1}, \dots, b_{i,|B_i|} \text{ for each } B_i = \{b_{i,1}, \dots, b_{i,|B_i|}\}?\end{aligned} \tag{3.10}$$

*Example* 22. For $n = 5$, $\sigma = 52314$ and $\mathcal{B} = \{\{1\}, \{2,3,5\}, \{4\}\}$, Query (3.9) corresponds to asking:

- what is the rank of element 1?
- what is the set of ranks of elements $\{2,3,5\}$?
- what is the rank of element 4?

and Query (3.10) corresponds to asking:

- what element is ranked at position 1?
- what is the set of elements ranked at positions $\{2,3,5\}$?
- what element is ranked at position 4?

Query (3.10) is certainly most natural when partition $\mathcal{B}$ is of the form $B_1 = \{1, \dots, n_1\}$, $B_2 = \{n_1 + 1, \dots, n_1 + n_2\}$, $\dots$, $B_r = \{\sum_{i=1}^{r-1} n_i + 1, \dots, n\}$. When applied to a deterministic full ranking $\sigma$, it then becomes: what are the $n_1$ first elements, $n_2$ second elements, $\dots$, $n_r$ last elements in $\sigma$? The answer to this query is the only partial ranking of the form $a_{1,1}, \dots a_{1,n_1} \succ \cdots \succ a_{r,1}, \dots, a_{r,n_r}$ that admits $\sigma$ as a linear extension. The answer to the analogous query about a random permutation $\Sigma$ drawn from a ranking model $p$ is thus naturally the marginal of $p$ on partial rankings of this form.

*Example* 23. For $k \in \{1, \dots, n-1\}$, the answer to Query (3.10) on a full ranking $\sigma$ for $\mathcal{B} = \{\{1\}, \dots, \{k\}, \{k+1, \dots, n\}\}$ is the ordered top-$k$ ranking $\sigma_1 \succ \cdots \succ \sigma_k \succ$ *the rest* induced by $\sigma$. Other example, for $n = 7$ and $\mathcal{B} = \{\{1,2,3\}, \{4,5\}, \{6,7\}\}$, the answer to Query (3.10) on a full ranking $\sigma \in \mathfrak{S}_7$ is the ranking $\sigma_1, \sigma_2 \succ \sigma_3, \sigma_4, \sigma_5 \succ \sigma_6, \sigma_7$.

We now give a rigorous definition for general absolute marginals. We call the shape of partition $\mathcal{B} = (B_1, \dots, B_r) \in \mathrm{Part}(\llbracket n \rrbracket)$ the tuple $\mathrm{shape}(\mathcal{B}) = (\lambda_1, \dots, \lambda_r)$ obtained by sorting the tuple $(|B_1|, \dots, |B_r|)$ in decreasing order. It is easy to see that the shape of a partition of $\llbracket n \rrbracket$ is a *partition of* $n$, a classic object in combinatorics, the definition of which is recalled here.

**Definition 24** (Partition of $n$)**.** A partition of $n$ is a tuple $\lambda = (\lambda_1, \dots, \lambda_r) \in \mathbb{N}^r$ such that $\lambda_1 \geq \cdots \geq \lambda_r \geq 1$ and $\sum_{i=1}^r \lambda_i = n$. The notation $\lambda \vdash n$ means that $\lambda$ is a partition of $n$.

For $\lambda = (\lambda_1, \dots, \lambda_r) \vdash n$, we define the set $\mathrm{Part}_\lambda(\llbracket n \rrbracket) = \{\mathcal{B} \in \mathrm{Part}(\llbracket n \rrbracket) \mid \mathrm{shape}(\mathcal{B}) = \lambda\}$. Equipped with these notations, we give a rigorous definition for absolute marginals.

**Definition 25** (Absolute marginals)**.** Let $\Sigma$ be a random permutation drawn from a ranking model $p$, $\mathcal{B} \in \mathrm{Part}(\llbracket n \rrbracket)$ be a partition of $\llbracket n \rrbracket$ and $\lambda = \mathrm{shape}(\mathcal{B})$.

- The direct marginal of $p$ on $\mathcal{B}$ is the law of $\Sigma(\mathcal{B})$ on $\mathrm{Part}_\lambda(\llbracket n \rrbracket)$, that is the collection of probabilities

$$\mathbb{P}[\Sigma(\mathcal{B}) = \mathcal{B}'] = \sum_{\sigma \in \mathfrak{S}_n,\ \sigma(\mathcal{B}) = \mathcal{B}'} p(\sigma) \qquad \text{for } \mathcal{B}' \in \mathrm{Part}_\lambda(\llbracket n \rrbracket).$$

- The reciprocal marginal of $p$ on $\mathcal{B}$ is the law of $\Sigma^{-1}(\mathcal{B})$ on $\mathrm{Part}_\lambda([\![n]\!])$, that is the collection of probabilities

$$\mathbb{P}[\Sigma^{-1}(\mathcal{B}) = \mathcal{B}'] = \sum_{\sigma\in\mathfrak{S}_n,\ \sigma^{-1}(\mathcal{B})=\mathcal{B}'} p(\sigma) \qquad \text{for } \mathcal{B}' \in \mathrm{Part}_\lambda([\![n]\!]).$$

Both are called absolute marginals *of shape* $\lambda$. Given an ordering of the elements of $\mathrm{Part}_\lambda([\![n]\!])$, we define the $|\mathrm{Part}_\lambda([\![n]\!])| \times |\mathrm{Part}_\lambda([\![n]\!])|$ square matrix

$$M_\lambda p = \Big(\mathbb{P}[\Sigma(\mathcal{B}) = \mathcal{B}']\Big)_{\mathcal{B},\mathcal{B}'\in\mathrm{Part}_\lambda([\![n]\!])}.$$

The matrix $M_\lambda p$ contains all the absolute marginals of $p$ of shape $\lambda$: the row indexed by $\mathcal{B}$ is the direct marginal of $p$ on $\mathcal{B}$ and the column indexed by $\mathcal{B}'$ is the reciprocal marginal of $p$ on $\mathcal{B}'$.

Definition 25 can seem intricate at first sight but it actually includes some simple and natural cases. Let us first consider the case $\lambda = (n-1, 1)$. Elements of $\mathrm{Part}_{(n-1,1)}([\![n]\!])$ are necessarily of the form $([\![n]\!] \setminus \{i\}, \{i\})$, with $i \in [\![n]\!]$. Then for $(i, j) \in [\![n]\!]^2$, we have the simplification

$$\mathbb{P}\Big[\Sigma([\![n]\!] \setminus \{i\}) = [\![n]\!] \setminus \{j\},\ \Sigma(\{i\}) = \{j\}\Big] = \mathbb{P}\left[\Sigma(i) = j\right].$$

The direct marginal of $p$ on $([\![n]\!] \setminus \{i\}, \{i\})$ is thus the probability distribution $(\mathbb{P}[\Sigma(i) = j])_{j\in[\![n]\!]}$ on $[\![n]\!]$, that is to say the law of the rank $\Sigma(i)$ of element $i$, and the reciprocal marginal of $p$ on $([\![n]\!] \setminus \{j\}, \{j\})$ is the probability distribution $(\mathbb{P}[\Sigma^{-1}(j) = i])_{i\in[\![n]\!]}$ on $[\![n]\!]$, that is to say the law of the element $\Sigma^{-1}(j)$ ranked in $j^{th}$ position. The matrix $M_{(n-1,1)}p$ is given by

$$M_{(n-1,1)}(p) = \begin{pmatrix} \mathbb{P}[\Sigma(1) = 1] & \cdots & \mathbb{P}[\Sigma(n) = 1] \\ \vdots & \ddots & \vdots \\ \mathbb{P}[\Sigma(1) = n] & \cdots & \mathbb{P}[\Sigma(n) = n] \end{pmatrix}.$$

The same reasoning shows that the matrix $M_\lambda p$ is given by

$$M_{(n-2,2)}p = \Big(\mathbb{P}[\Sigma(\{i, i'\}) = \{j, j'\}]\Big)_{\substack{1\le i<i'\le n\\ 1\le j<j'\le n}} \qquad \text{for } \lambda = (n-2, 2),$$

$$\text{and} \qquad M_{(n-2,1,1)}p = \Big(\mathbb{P}[\Sigma(i) = j, \Sigma(i') = j']\Big)_{\substack{1\le i\ne i'\le n\\ 1\le j\ne j'\le n}} \qquad \text{for } \lambda = (n-2, 1, 1).$$

More generally, for $\lambda = (\lambda_1, \ldots, \lambda_r) \vdash n$ and $\mathcal{B} = (B_1, \ldots, B_r) \in \mathrm{Part}_\lambda([\![n]\!])$, the partition $\sigma(\mathcal{B})$ is entirely characterized by $(\sigma(B_1), \ldots, \sigma(B_{r-1}))$. A marginal of shape $\lambda$ is thus the answer to a query that concerns the ranks of $n - \lambda_1$ elements (for a direct marginal) or the elements that are ranked at $n - \lambda_1$ positions (for a reciprocal marginal).

### 3.2.3 Absolute versus Relative Marginals

Here and throughout the rest of this section, we call the marginals $M_A p$ of a ranking model $p$ *relative marginals* to contrast with *absolute marginals*. The justification for these terms is the following. For $A = \{a_1, \ldots, a_k\} \in \mathcal{P}([\![n]\!])$, the query to which $M_A p$ is the answer is "what is the *relative* order of elements $a_1, \ldots, a_k$ in a random full ranking $\Sigma$ drawn from $p$?" The marginal $M_A p$ thus contains the part of information related to the ranks of elements $a_1, \ldots, a_k$ in the

ranking $\Sigma_{|A}$. These ranks are relative: they only characterize the relative position of an element of $A$ compared to the others.

By contrast, absolute marginals contain parts of information related to the ranks of elements in the full ranking $\Sigma$, that is to say their *absolute* ranks. Let $\lambda \vdash n$ and $\mathcal{B} \in \mathrm{Part}_\lambda([\![n]\!])$. The direct marginal of $p$ on $\mathcal{B}$ is the law of $\Sigma(\mathcal{B})$. It thus contains information about the absolute ranks of the elements of each $B_i$. The reciprocal marginal of $p$ on $\mathcal{B}$ is the law of $\Sigma^{-1}(\mathcal{B})$. It contains information about the elements that are placed at the absolute positions specified by each $B_i$. Such information thus also has an absolute nature.

The previous explanation is quite intuitive but not very formal. We now provide a rigorous interpretation using group actions and translations: we show that under a relabeling of the elements of $[\![n]\!]$, absolute marginals are stable but relative marginals are not. A relabeling of the elements of $[\![n]\!]$ is naturally defined as the action of a permutation (we refer the reader to Fulton and Harris, 1991, for background on group theory).

**Definition 26** (Action of $\mathfrak{S}_n$ over $\Gamma_n$ and $\mathrm{Part}([\![n]\!])$)**.** Let $\tau \in \mathfrak{S}_n$ be a permutation.

- The action of $\tau$ on an incomplete ranking $\pi = \pi_1 \dots \pi_k \in \Gamma_n$ is defined by $\tau(\pi) = \tau(\pi_1) \dots \tau(\pi_k)$.
- The action of $\tau$ on a partition $\mathcal{B} \in \mathrm{Part}([\![n]\!])$ of $[\![n]\!]$ is defined by $\tau(\mathcal{B})$.

It is easy to see that the mappings $\pi \mapsto \tau(\pi)$ and $\mathcal{B} \mapsto \tau(\mathcal{B})$ are both actions of $\mathfrak{S}_n$, respectively on $\Gamma_n$ and $\mathrm{Part}([\![n]\!])$. The proof is left to the reader.

*Example* 27. Let $\tau \in \mathfrak{S}_5$ be the permutation defined by $\tau(1) = 3, \tau(2) = 1, \tau(3) = 5, \tau(4) = 4, \tau(5) = 2$. Then for $\pi = 3421$ and $\mathcal{B} = \{\{1,3\},\{2,4\},\{5\}\}$, one has

$$\tau(\pi) = 5413 \qquad \text{and} \qquad \tau(\mathcal{B}) = \{\{3,5\},\{1,4\},\{2\}\}.$$

*Remark* 28 (Right action of $\mathfrak{S}_n$). The action of a permutation $\tau$ on a full ranking $\sigma = \sigma_1 \dots \sigma_n \in \mathfrak{S}_n$ is by Definition 26 equal to $\tau(\sigma_1) \dots \tau(\sigma_n)$. We point out that if $\sigma$ is seen as a permutation with $\sigma_i \equiv \sigma^{-1}(i)$ for each $i \in [\![n]\!]$ then the permutation associated to the ranking $\tau(\sigma)$ is the permutation $\sigma'$ such that $\sigma'^{-1}(i) = \tau(\sigma^{-1}(i))$, namely $\sigma' = \sigma\tau^{-1}$. The action $\sigma \mapsto \tau(\sigma)$ on $\mathfrak{S}_n$ is thus equivalent to the right action of $\mathfrak{S}_n$ on itself.

Having defined the action of $\mathfrak{S}_n$ on $\Gamma_n$ and $\mathrm{Part}([\![n]\!])$ we can now define the associated translation operators on the spaces $L(\Gamma_n)$ and $L(\mathrm{Part}([\![n]\!]))$.

**Definition 29** (Translation operators)**.** Let $\tau \in \mathfrak{S}_n$.

- We define the translation operator $T_\tau$ on $L(\Gamma_n)$ by $T_\tau \delta_\pi = \delta_{\tau(\pi)}$ for all $\pi \in \Gamma_n$ or equivalently by
$$T_\tau F(\pi) = F(\tau^{-1}(\pi)) \qquad \text{for all } F \in L(\Gamma_n) \text{ and } \pi \in \Gamma_n.$$
- We define the translation operator $\mathcal{T}_\tau$ on $L(\mathrm{Part}([\![n]\!]))$ by $\mathcal{T}_\tau \delta_\mathcal{B} = \delta_{\tau(\mathcal{B})}$ for all $\mathcal{B} \in \mathrm{Part}([\![n]\!])$ or equivalently by
$$\mathcal{T}_\tau f(\mathcal{B}) = f(\tau^{-1}(\mathcal{B})) \qquad \text{for all } f \in L(\mathrm{Part}([\![n]\!])) \text{ and } \mathcal{B} \in \mathrm{Part}([\![n]\!]).$$

The translation under a permutation $\tau \in \mathfrak{S}_n$ of a function $F \in L(\Gamma_n)$ or a function $f \in L(\mathrm{Part}([\![n]\!]))$ is the function obtained after relabeling the elements of $[\![n]\!]$ according to $\tau$.

*Example* 30. With the same permutation $\tau \in \mathfrak{S}_n$ as in Example 27, one has for instance

$$T_\tau\left(0.2\delta_{3421} + 0.3\delta_{2413} + 0.5\delta_{4312}\right) = 0.2\delta_{5413} + 0.3\delta_{1435} + 0.5\delta_{4531}$$
$$\mathcal{T}_\tau\left(0.6\delta_{\{\{1,3\},\{2,4\},\{5\}\}} + 0.4\delta_{\{\{1,3,4\},\{2\},\{5\}\}}\right) = 0.6\delta_{\{\{3,5\},\{1,4\},\{2\}\}} + 0.4\delta_{\{\{3,5,4\},\{1\},\{2\}\}}.$$

Equipped with these definitions, we can now state the different stability properties of the marginals, justifying even more the names "absolute" and "relative" marginals.

**Proposition 31** (Stability / Instability under relabeling)**.** *Let $p$ be a ranking model and $\tau \in \mathfrak{S}_n$.*

- *For $A \in \mathcal{P}(\llbracket n \rrbracket)$,*

$$\tau(\sigma)_{|A} = \tau\left(\sigma_{|\tau^{-1}(A)}\right) \quad \text{for all } \sigma \in \mathfrak{S}_n \qquad \text{and thus} \qquad M_A T_\tau p = T_\tau M_{\tau^{-1}(A)} p.$$

- *For $\mathcal{B}' \in \mathrm{Part}(\llbracket n \rrbracket)$,*

$$(\tau(\sigma))^{-1}(\mathcal{B}') = \tau\left(\sigma^{-1}(\mathcal{B}')\right) \quad \text{for all } \sigma \in \mathfrak{S}_n \qquad \text{and thus} \qquad M_{\mathcal{B}'} T_\tau p = \mathcal{T}_\tau M_{\mathcal{B}'} p,$$

*where $M_{\mathcal{B}'}$ is the operator associated to reciprocal marginals.*

*Proof.* Let $\sigma \in \mathfrak{S}_n$. By definition

$$\tau(\sigma)_{|A} = (\tau(\sigma_1) \dots \tau(\sigma_n))_{|A} = \tau(\sigma_{i_1}) \dots \tau(\sigma_{i_{|A|}})$$

where $1 \leq i_1 \leq \dots \leq i_{|A|} \leq n$ are all the indexes such that $\tau(\sigma_{i_j}) \in A$ or equivalently such that $\sigma_{i_j} \in \tau^{-1}(A)$. In other words, $\sigma_{i_1} \dots \sigma_{i_{|A|}} = \sigma_{|\tau^{-1}(A)}$. Hence $\tau(\sigma)_{|A} = \tau(\sigma_{|\tau^{-1}(A)})$ and the extension to marginals is immediate. For the second property we recall that $\tau(\sigma)$ is equal to the permutation $\sigma\tau^{-1}$, so that

$$(\tau(\sigma))^{-1}(\mathcal{B}') = (\sigma\tau^{-1})^{-1}(\mathcal{B}') = (\tau\sigma^{-1})(\mathcal{B}') = \tau\left(\sigma^{-1}(\mathcal{B}')\right).$$

Again, the extension to marginals is immediate. □

Proposition 31 shows that absolute reciprocal marginals are stable under relabeling. For $\mathcal{B}' \in \mathrm{Part}(\llbracket n \rrbracket)$ indeed, relabeling the elements of $\llbracket n \rrbracket$ and then computing the reciprocal marginal gives the same result as doing it in the other way round. By contrast, the relative marginal of $p$ on a subset $A \in \mathcal{P}(\llbracket n \rrbracket)$ is stable under the action of a permutation $\tau \in \mathfrak{S}_n$ only if $\tau(A) = A$. These properties are an additional justification for the terms "absolute" and "relative" marginals. The former are stable under relabeling of the elements, hence they carry global or absolute rank information. The latter are highly sensitive to relabeling of the elements: they carry local or relative rank information.

### 3.2.4 Rank information localization

We now come back to the question: how much information does a marginal of a ranking model $p$, or more generally a collection of marginals of $p$, contain about $p$? We address it for absolute or relative marginals.

In particular one would like to know if the knowledge of a certain collection of marginals implies the knowledge of $p$. If $p = \delta_{\sigma^*}$ is the Dirac distribution on a permutation $\sigma^* \in \mathfrak{S}_n$, then marginals of $p$ are the answers to queries on $\sigma^*$ and it is well known that the exact knowledge of $\sigma^*$ can be recovered from a collection of simple queries. For example, the knowledge of $\sigma^*(1), \dots, \sigma^*(n-1)$, the answers to $n-1$ $n$-ary queries, is sufficient to characterize $\sigma^*$. Other example, the knowledge of $(\sigma^*_{|\{a,b\}})_{1 \leq a < b \leq n}$, the answers of $n(n-1)/2$ binary queries, is sufficient to characterize $\sigma^*$.[2]

[2]It is well known that one actually only needs $O(n \log n)$ binary queries to characterize a full ranking, see for instance Knuth (1973).

This is however not true for a general ranking model $p$. A simple dimensional argument can show it. As a probability distribution over a set of $n!$ elements, $p$ is a vector characterized by $n! - 1$ parameters. Each pairwise marginal $M_{\{a,b\}}p$ of $p$ being a probability distribution over a set of 2 elements, it is characterized by 1 parameter. Thus the knowledge of all the pairwise marginals of $p$ enables to characterize at most $n(n-1)/2$ parameters of $p$, not enough to freeze the $n! - 1$ degrees of freedom.

This argument only provides part of the answer to the question. If one observes a collection of marginals of $p$, it may be indeed that the sum of the number of parameters characterized by each marginal exceeds $n! - 1$. The parameters characterized by different marginals may however not be linearly independent so that this situation would not necessarily mean that $p$ is entirely characterized.

The relevant question that one should turn to is therefore: given a family of marginals of $p$, how much more information about $p$ does a marginal contain? To answer this question, one must exploit the natural structure of absolute or relative marginals. The description of the structure of absolute marginals first requires the introduction of the dominance order, a classic structure in combinatorics (see for instance Stanley, 1986).

**Definition 32** (Dominance order on partitions of $n$)**.** For two partitions $\lambda = (\lambda_1, \dots, \lambda_r), \mu = (\mu_1, \dots, \mu_s) \vdash n$, we say that $\lambda$ is dominated by $\mu$ and write $\lambda \trianglelefteq \mu$ if $s \leq r$ and $\sum_{i=1}^{j} \lambda_i \leq \sum_{i=1}^{j} \mu_i$ for all $j \in \{1, \dots, s\}$. It is easy to see that $\trianglelefteq$ is a partial order over the set of partitions of $[\![n]\!]$ and we denote by $\triangleleft$ its associated strict partial order.

The top of the dominance order is: $(n) \trianglerighteq (n-1,1) \trianglerighteq (n-2,2) \trianglerighteq (n-2,1,1) \trianglerighteq \dots$ The Hasse diagram of the dominance order is represented on Figure 3.3 for $n = 6$ (the reversed order is actually represented, for reasons detailed below). The structure of absolute and relative marginals is characterized in the following proposition, the proof of which is left in Appendix.

**Proposition 33** (Structure of absolute and relative marginals)**.** *Let $p$ be a ranking model and $\Sigma$ be a random permutation drawn from $p$.*

1. *For $A, B \in \mathcal{P}([\![n]\!])$ with $B \subset A$, one has for all $\pi \in \Gamma(B)$*

$$M_B p(\pi) = \sum_{\pi' \in \Gamma(A),\ \pi'_{|A} = \pi} M_A p(\pi')$$

2. *For $\lambda, \mu \vdash n$ with $\mu \trianglerighteq \lambda$, there exists a linear operator $\mathcal{M}_{\lambda,\mu} : \mathbb{R}^{|\operatorname{Part}_\lambda([\![n]\!])| \times |\operatorname{Part}_\lambda([\![n]\!])|} \to \mathbb{R}^{|\operatorname{Part}_\mu([\![n]\!])| \times |\operatorname{Part}_\mu([\![n]\!])|}$ such that*

$$M_\mu p = \mathcal{M}_{\lambda,\mu} M_\lambda p.$$

Proposition 33 establishes some structural relationships between the relative marginals of a ranking model $p$ or between its absolute marginals. For relative marginals, it implies that the knowledge of the marginal $M_A p$ of $p$ on a subset $A \in \mathcal{P}([\![n]\!])$ induces the knowledge of the marginal $M_B p$ on any subset $B \in \mathcal{P}(A)$. For absolute marginals, it says that the knowledge of the matrix $M_\lambda p$ of all the marginals of $p$ of shape $\lambda \vdash n$ induces the knowledge of the matrix of all the marginals $M_\mu p$ for $\mu \trianglerighteq \lambda$.

In both cases, when the knowledge of a marginal induces the knowledge of another, it is natural to say that the first one carries more information. The question is then: how much more information? or more precisely what additional part of information does it carry? In other words, from the knowledge of the second marginal, what additional part of information is needed to recover the first one?

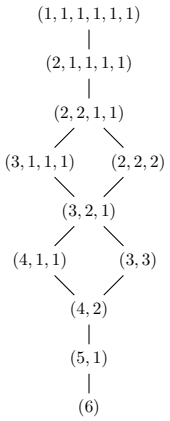



Figure 3.3: Reversed Hasse diagram of the dominance order on partitions of $n$ for $n = 6$

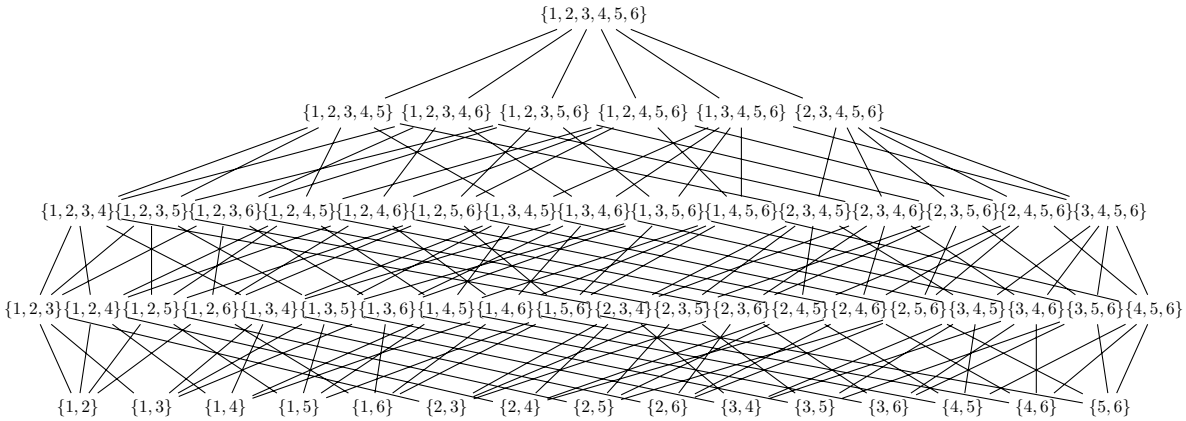



Figure 3.4: Hasse diagram of the elements of $\mathcal{P}([\![n]\!])$ for the partial order defined by the subset inclusion, for $n = 6$

To quantify this part of information, one needs to finely exploit the structure of the marginals and their associated partial order, namely the dominance order on partitions of $n$ for absolute marginals and the partial order defined by subset inclusion on $\mathcal{P}([\![n]\!])$ for relative marginals. We provide a representation of their Hasse diagrams on Figures 3.3 and 3.4. On Figure 3.3 we represent the inverse dominance order so that partitions associated to marginals with more information appear higher.

We show in Subsection 3.2.5 that Fourier analysis on the symmetric group is exactly suited to quantify the additional part of information between absolute marginals, but not between relative marginals. This is the purpose of the MRA representation introduced in this thesis.

### 3.2.5 Fourier analysis localizes absolute rank information but not relative rank information

Abstract Fourier analysis consists in representing functions as sums of projections onto spaces that are invariant under translations (see Diaconis, 1989). By Proposition 31, absolute marginals are stable under translation operators $\mathcal{T}_\tau$ for $\tau \in \mathfrak{S}_n$. It is therefore natural to expect that Fourier analysis localizes absolute rank information. Though it was already explained in details in the literature (see for instance Diaconis, 1988; Huang et al., 2009a), we develop this interpretation here with a slightly different point of view to fit with the general approach of this thesis.

A transitive action[3] $(g, x) \mapsto g \cdot x$ of a group $G$ on a finite set $\mathcal{X}$ naturally defines a family of translation operators $T_g$ on $L(\mathcal{X})$ by $T_g \delta_x = \delta_{g \cdot x}$ or equivalently by $T_g f(x) = f(g^{-1} \cdot x)$ for any $f \in L(\mathcal{X})$ and $x \in \mathcal{X}$. As a permutation matrix, $T_g$ is diagonalizable for each $g \in G$. If $G$ is commutative then the translation operators two-by-two commute and a classic result from linear algebra states that there exists a basis of $L(\mathcal{X})$ in which all the translation operators are diagonal. This basis is a Fourier basis (see for instance Diaconis, 1989).

When $G$ is not commutative, this result does not hold anymore (if there existed a basis where all the translation operators are diagonal then they would commute to-by-two and $G$ would be commutative). One however has $T_{gg'} = T_g T_{g'}$ for any $g, g' \in G$. This means that the mapping $g \mapsto T_g$ is a representation of $G$ on $L(\mathcal{X})$. A classic result from group representation theory then says that the following decomposition holds

$$L(\mathcal{X}) \cong \bigoplus_\rho m_\rho V_\rho, \tag{3.11}$$

where each $\rho$ is an irreducible representation of $G$, $V_\rho$ its associated linear space and $m_\rho$ a nonnegative integer. Here and throughout the rest of the thesis, the symbol $\cong$ in (3.11) means that the two spaces are isomorphic as representations (of the group $G$ in the present case). In the particular case where $\mathcal{X} = G$, one has $m_\rho = \dim V_\rho$ for each irreducible representation $\rho$. Equation (3.11) means in practice that there exists an isomorphism of representations $\Phi : \bigoplus_\rho m_\rho V_\rho \to L(\mathcal{X})$ such that any function $f \in L(\mathcal{X})$ admits a unique decomposition

$$f = \Phi \sum_\rho P_\rho f \qquad \text{with } P_\rho f \in m_\rho V_\rho \text{ for each irreducible representation } \rho. \tag{3.12}$$

For each $\rho$, $P_\rho f$ can be seen as a "projection" of $f$ onto $m_\rho V_\rho$, which thus localizes a certain part of information about $f$ that is invariant under translations. We develop this interpretation for the symmetric group.

Representations of the symmetric group have been thoroughly studied in the literature (see for instance James and Kerber, 1981; Ceccherini-Silberstein et al., 2010; Sagan, 2013). Each irreducible representation of $\mathfrak{S}_n$ is indexed by a partition of $n$ (see Definition 21). The spaces of the irreducible representations are called the Specht modules. They are denoted by $S^\lambda$ and their dimensions by $d_\lambda$ for $\lambda \vdash n$. Decomposition (3.11) for $\mathfrak{S}_n$ then gives the following result.

**Proposition 34** (Fourier decomposition of $L(\mathfrak{S}_n)$)**.** *The following decomposition holds*

$$L(\mathfrak{S}_n) \cong \bigoplus_{\lambda \vdash n} d_\lambda S^\lambda \qquad \textit{with } d_\lambda = \dim S^\lambda \textit{ for each } \lambda \vdash n.$$

The best way to formulate properly Equation (3.12) for the symmetric group, is to use the Fourier transform. In practice, each space $d_\lambda S^\lambda$ in the decomposition of Proposition 34 is

[3] an action $(g, x) \mapsto g \cdot x$ of a group $G$ on a finite set $\mathcal{X}$ is transitive if it has only one orbit or equivalently if for any $x, x' \in \mathcal{X}$, there exists $g \in G$ such that $g \cdot x = x'$

replaced by the isomorphic space $\mathbb{R}^{d_\lambda \times d_\lambda}$ of $d_\lambda$-square matrices. Let $\rho_\lambda$ be a representative of the irreducible representation indexed by $\lambda \vdash n$. By definition, the latter is a mapping $\rho_\lambda : G \to \mathbb{R}^{d_\lambda \times d_\lambda}$ such that $\rho_\lambda(id) = I_{d_\lambda}$, the identity matrix of size $d_\lambda$, and $\rho_\lambda(\tau\tau') = \rho_\lambda(\tau)\rho_\lambda(\tau')$. In particular $\rho_\lambda(\tau)$ is invertible for any $\tau \in \mathfrak{S}_n$ with $\rho_\lambda(\tau)^{-1} = \rho_\lambda(\tau^{-1})$. One can also choose $\rho_\lambda$ such that $\rho_\lambda(\tau)$ is an orthogonal matrix for all $\tau \in \mathfrak{S}_n$, a classic example being the Young Orthogonal Representation (see for example James and Kerber, 1981).

**Definition 35** (Fourier transform)**.** Let $f \in L(\mathfrak{S}_n)$. For each $\lambda \vdash n$, the Fourier coefficient of $f$ indexed by $\lambda$ is the $d_\lambda \times d_\lambda$ matrix defined by

$$\widehat{f}(\lambda) = \sum_{\tau \in \mathfrak{S}_n} f(\tau)\rho_\lambda(\tau).$$

The Fourier transform is the mapping $\mathcal{F} : L(\mathfrak{S}_n) \to \bigoplus_{\lambda \vdash n} \mathbb{R}^{d_\lambda \times d_\lambda}, f \mapsto (\widehat{f}(\lambda))_{\lambda \vdash n}$.

A classic result in group representation theory is that the Fourier transform is an isometry. In particular it is invertible, the inverse Fourier transform being given by the following formula.

**Proposition 36** (Inverse Fourier transform)**.** *Let $f \in L(\mathfrak{S}_n)$. For any $\sigma \in \mathfrak{S}_n$, one has*

$$f(\sigma) = \frac{1}{n!} \sum_{\lambda \vdash n} d_\lambda \operatorname{tr}\left[\rho_\lambda(\sigma)^\top \widehat{f}(\lambda)\right],$$

*where for any matrix $M$, $M^\top$ denotes its transpose and $\operatorname{tr}(M)$ its trace.*

Proposition 36 is a statement of Equation (3.12) for functions of $\mathfrak{S}_n$. More precisely, the correspondence is made with $P_{\rho_\lambda} f := \widehat{f}(\lambda)$ for any $\lambda \vdash n$ and $\Phi : \bigoplus_{\lambda \vdash n} \mathbb{R}^{d_\lambda \times d_\lambda} \to L(\mathfrak{S}_n)$ the operator defined by $\Phi((F_\lambda)_{\lambda \vdash n})(\sigma) = (1/n!) \sum_{\lambda \vdash n} d_\lambda \operatorname{tr}\left[\rho_\lambda(\sigma)^\top F_\lambda\right]$ for all $\sigma \in \mathfrak{S}_n$.

Propositions 34 and 36 are the statements of Equations (3.11) and (3.12) for functions of $\mathfrak{S}_n$. To show that the Fourier transform localizes absolute rank information, we also need to state them for functions of $\mathrm{Part}([\![n]\!])$. From now on, we fix a partition $\lambda \vdash n$. It is easy to see that the action of $\mathfrak{S}_n$ on $\mathrm{Part}([\![n]\!])$ from Definition 26 is transitive on each subset $\mathrm{Part}_\lambda([\![n]\!])$. The translation operators $\mathcal{T}_\tau$ from Definition 29 can thus all be restricted to $L(\mathrm{Part}_\lambda([\![n]\!]))$ and the mapping $\tau \mapsto \mathcal{T}_\tau$ defines a representation of $\mathfrak{S}_n$ on the latter. Its isomorphism class is called a Young module and denoted by $M^\lambda$ in the literature. The equivalent of Decomposition (3.11) for $M^\lambda$ is given by Young's rule, a classic result in the representation theory of the symmetric group (see for instance James and Kerber, 1981).

**Proposition 37** (Young's rule)**.** *For $\lambda \vdash n$*

$$M^\lambda \cong \bigoplus_{\mu \trianglerighteq \lambda} K_{\mu,\lambda} S^\mu,$$

*where the $K_{\mu,\lambda}$'s are nonnegative integers called the Kotska's numbers, with $K_{\lambda,\lambda} = 1$.*

*Example* 38 (Decomposition of some Young modules). Proposition 37 provides the following decompositions

$$\begin{aligned}
M^{(n)} &\cong S^{(n)} \\
M^{(n-1)} &\cong S^{(n-1,1)} \oplus S^{(n)} \\
M^{(n-2,2)} &\cong S^{(n-2,2)} \oplus S^{(n-1,1)} \oplus S^{(n)} \\
M^{(n-2,1,1)} &\cong S^{(n-2,1,1)} \oplus S^{(n-2,2)} \oplus 2S^{(n-1,1)} \oplus S^{(n)}.
\end{aligned}$$

Obtaining an explicit equivalent of (3.12) for functions of $M^\lambda$ is tedious and not necessary for our purpose. We only use a part of *James' Submodule Theorem* (see Huang et al., 2009a). For $m$ square matrices $F_1, \dots, F_m$ of respective sizes $d_1, \dots, d_m$ we define the $(d_1 + \cdots + d_m)$-square matrix

$$\bigoplus_{i=1}^{m} F_i = \begin{pmatrix} F_1 & & \\ & \ddots & \\ & & F_m \end{pmatrix}.$$

**Theorem 39** (James' Submodule Theorem)**.** *There exists an orthogonal matrix $C_\lambda$ such that for all $f \in L(\mathfrak{S}_n)$,*

$$M_\lambda f = C_\lambda \left[ \bigoplus_{\mu \trianglerighteq \lambda} \bigoplus_{l=1}^{K_{\mu,\lambda}} \widehat{f}(\mu) \right] C_\lambda^\top.$$

The correspondence with (3.12) is obtained with $P_{\rho_\mu} M_\lambda f = \widehat{f}(\mu)$ for any $\mu \trianglerighteq \lambda$ and $f \in L(\mathfrak{S}_n)$ and $\Phi_\lambda((F_\mu)_{\mu \trianglerighteq \lambda}) = C_\lambda[\bigoplus_{\mu \triangleright \lambda} \bigoplus_{l=1}^{K_{\mu,\lambda}} F_\mu] C_\lambda^\top$. As $K_{\lambda,\lambda} = 1$ one thus has

$$M_\lambda f = \Phi_\lambda((\widehat{f}(\mu))_{\mu \trianglerighteq \lambda}) \qquad \text{or equivalently} \qquad \Phi_\lambda \widehat{f}(\lambda) = M_\lambda f - \Phi_\lambda((\widehat{f}(\mu))_{\mu \triangleright \lambda}). \tag{3.13}$$

Equation (3.13) shows first that if one knows the Fourier coefficients $\widehat{f}(\mu)$ for all $\mu \trianglerighteq \lambda$ then he knows $M_\lambda f$, that is to say all the absolute marginals of $f$ of shape $\lambda$. In other words the part of information related to absolute marginals of shape $\lambda$ is divided over the Fourier coefficients of partitions $\mu \trianglerighteq \lambda$. The second formula of Equation (3.13) then shows that if one knows all the Fourier coefficients of $f$ for partitions $\mu \triangleright \lambda$, or equivalently all the absolute marginals of $f$ of shape $\mu \triangleright \lambda$, the missing part of information to know the absolute marginals of shape $\lambda$ is contained in $\widehat{f}(\lambda)$. The Fourier coefficient $\widehat{f}(\lambda)$ thus localizes the part of information of $f$ specific to its absolute marginals of shape $\lambda$.

At last, the part of information carried by each Fourier coefficient being by construction invariant under translations, it cannot be specific to a relative marginal. It is the purpose of the MRA representation to localize specific information of relative marginals.

# Part II

# Contributions

# Chapter 4

# Multiresolution analysis of incomplete rankings

This chapter introduces the multiresolution analysis (MRA) of incomplete rankings, the major theoretical contribution of this thesis. First, Section 4.1 establishes the fundamental result about the multiresolution decomposition of any space $L(\Gamma(A))$ for $A \in \mathcal{P}(\llbracket n \rrbracket)$. This result is used in Section 4.2 to construct the MRA representation and the wavelet transform. Section 4.3 treats the associated computational aspects. At last, Section 4.4 introduces a wavelet basis consistent with the multiresolution decomposition.

## Contents

## 4.1 Multiresolution decomposition

In this section we introduce the main objects of the multiresolution analysis of incomplete rankings and establish the central result about the multiresolution decomposition. We recall that the null space of any operator $T$ is denoted by $\ker T$ and for any finite set $E$, we set $\bar{\mathcal{P}}(E) := \mathcal{P}(E) \cup \{\emptyset\}$. We denote by convention $\bar{0}$ the unique injective word of content $\emptyset$ and length 0, and set $\bar{\Gamma}_n := \Gamma_n \cup \{\bar{0}\}$. We extend naturally the marginal operators to the space $L(\bar{\Gamma}_n)$ and define by convention the marginal operator on $\emptyset$ by $M_\emptyset : L(\bar{\Gamma}_n) \to L(\Gamma(\bar{0})),\ F \mapsto (\sum_{\pi \in \bar{\Gamma}_n} F(\pi))\delta_{\bar{0}}$.

### 4.1.1 Main definitions

The multiresolution analysis (MRA) of incomplete rankings is built on two central objects: the spaces $H_B$ and the operators $\phi_A$, for $A, B \in \bar{\mathcal{P}}([\![n]\!])$.

**Definition 40** (Spaces $H_B$)**.** We set $H_\emptyset = \mathbb{R}\bar{0} = L(\Gamma(\bar{0}))$ and define for $B \in \mathcal{P}([\![n]\!])$ the linear space

$$H_B = \{F \in L(\Gamma(B)) \mid M_{B'}F = 0 \text{ for all } B' \subsetneq B\} = L(\Gamma(B)) \cap \bigcap_{B' \subsetneq B} \ker M_{B'}.$$

As we shall show thereafter, $H_B$ is the space where "live" the components that localize specific information of marginals on $B$, for $B \in \bar{\mathcal{P}}([\![n]\!])$. It can already be seen at first sight as a good candidate for this purpose since two functions $F$ and $G$ in $L(\Gamma(B))$ have the same marginals on all strict subsets of $B$ if and only if $F - G \in H_B$. Thus the projection of $F$ onto $H_B$ (in parallel to any space supplementary to $H_B$) contains information about $F$ that is specific to $B$. The results in this section will show that for any $A \in \mathcal{P}([\![n]\!])$, the structure of the space $L(\Gamma(A))$ is somehow equivalent to that of the sum of spaces $\bigoplus_{B \in \bar{\mathcal{P}}(A)} H_B$.

*Example* 41. In this example, we denote by $F_\pi$ the value of a function $F \in L(\bar{\Gamma}_n)$ on a ranking $\pi \in \bar{\Gamma}_n$. For $B = \{1, 2\}$, the space $H_{\{1,2\}}$ is simply equal to the one-dimensional space

$$H_{\{1,2\}} = \{(F_{12}, F_{21}) \in \mathbb{R}^2 \mid F_{12} + F_{21} = 0\}.$$

For $B = \{1, 2, 3\}$, $H_{\{1,2,3\}}$ is the space of vectors $(F_{123}, F_{132}, F_{213}, F_{231}, F_{312}, F_{321}) \in \mathbb{R}^6$ that satisfy the system

$$\begin{cases} F_{123} + F_{132} + F_{213} + F_{231} + F_{312} + F_{321} = 0 \\ F_{123} + F_{132} + \quad + \quad + F_{312} + \quad = 0 \\ F_{123} + F_{132} + F_{213} + \quad + \quad = 0 \\ F_{123} + \quad + F_{213} + F_{231} + \quad + \quad = 0 \end{cases}$$

It is easy to show that $H_{\{1,2,3\}}$ has dimension 2. Calculating the dimension of $H_B$ for any $B \in \mathcal{P}([\![n]\!])$ is not straightforward however. The result is given by Theorem 51.

The other central objects of the MRA are the operators $\phi_A$ for $A \in \bar{\mathcal{P}}([\![n]\!])$. Their definition relies on the following concept.

**Definition 42** (Contiguous subword)**.** Word $\pi' \in \Gamma_n$ is a *contiguous subword* of word $\pi \in \Gamma_n$ if there exists $i \in \{1, \dots, |\pi| - |\pi'| + 1\}$ such that $\pi' = \pi_i \pi_{i+1} \dots \pi_{i+|\pi'|-1}$. This is denoted by $\pi' \sqsubset \pi$.

*Example* 43. The contiguous subwords of $\pi = 2314$ are 23, 31, 14, 231, 314 and 2314. Notice that 214 is a subword of $\pi$ but not a contiguous subword.

**Definition 44** (Operators $\phi_A$)**.** For $A \in \bar{\mathcal{P}}([\![n]\!])$, we define the linear operator $\phi_A : L(\bar{\Gamma}_n) \to L(\Gamma(A))$ on the Dirac function of a ranking $\pi \in \bar{\Gamma}_n$ by

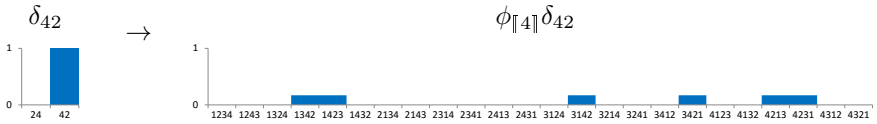
$$\phi_A \delta_{\bar{0}} = \frac{1}{|A|!} \mathbb{1}_{\Gamma(A)} \qquad \text{and} \qquad \phi_A \delta_\pi = \frac{1}{(|A| - |\pi| + 1)!} \mathbb{1}_{\{\sigma \in \Gamma(A) \mid \pi \sqsubset \sigma\}} \quad \text{if} \quad \pi \neq \bar{0}.$$

Though the operator $\phi_A$ for $A \in \bar{\mathcal{P}}([\![n]\!])$ is defined as a mapping from $L(\bar{\Gamma}_n)$ to $L(\Gamma(A))$, we shall see it as an "embedding operator". Indeed one has by Definition 44, $\phi_A F = 0$ for any $F \in L(B)$ with $B \not\subset A$. This means that $\phi_A$ maps $\bigoplus_{B \notin \bar{\mathcal{P}}(A)} L(\Gamma(B))$ to 0 and that it is only relevant to consider its effect on $\bigoplus_{B \in \bar{\mathcal{P}}(A)} L(\Gamma(B))$. In practice, it will be used to embed $\bigoplus_{B \in \bar{\mathcal{P}}(A)} H_B$ into $L(\Gamma(A))$.

*Example* 45. For $A = \{1, 2, 3, 4\}$ and $\pi = 24$ one has

$$\phi_{\{1,2,3,4\}} \delta_{24} = \frac{1}{6} \left( \delta_{1324} + \delta_{3124} + \delta_{1243} + \delta_{3241} + \delta_{2413} + \delta_{2431} \right).$$

This is represented by the following graphics.




Definition 44 is certainly not intuitive for an embedding operator. A more natural definition would certainly be to send the Dirac function of a ranking $\pi$ to the (normalized) sum of the Dirac functions of all the rankings that admit $\pi$ as a subword, not just as a contiguous subword. A detailed comparison with this operator and explanation for that choice is provided in Subsection 4.1.3.

The normalization coefficient in Definition 44 comes from the following Lemma. Its normalization purpose will however become clear in the proof of Lemma 49.

**Lemma 46.** *For $A \in \mathcal{P}([\![n]\!])$ and $\pi \in \Gamma_n$ with $c(\pi) \subset A$, the number of rankings in $\Gamma(A)$ that admit $\pi$ as a contiguous subword is equal to*

$$|\{\sigma \in \Gamma(A) \mid \pi \sqsubset \sigma\}| = (|A| - |\pi| + 1)!.$$

*Proof.* A ranking $\sigma \in \Gamma(A)$ such that $\pi \sqsubset \sigma$ is of the form $\sigma = a_1 \dots a_i \pi a_{i+1} \dots a_k$ with $k = |A| - |\pi|$. It can thus be seen as a linear order over the set $\{a_1, \dots, a_k\} \cup \{\pi\}$. As there are $(k+1)!$ such linear orders, this concludes the proof. □

### 4.1.2 Main result

The following theorem exploits the properties of both the spaces $H_B$ and operators $\phi_A$. It is the base of the entire MRA of incomplete rankings.

**Theorem 47** (Multiresolution decomposition)**.** *For any $A \in \bar{\mathcal{P}}([\![n]\!])$, one has the decomposition*

$$L(\Gamma(A)) = \bigoplus_{B \in \bar{\mathcal{P}}(A)} \phi_A \left( H_B \right).$$

*In addition, for $B \in \bar{\mathcal{P}}(A)$,*

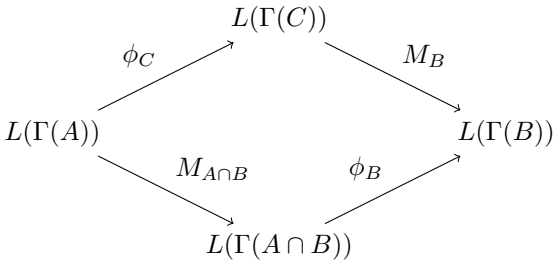
1. *$\phi_A$ is injective on $H_B$:* $\ker \phi_A \cap H_B = \{0\}$,
2. *for all $F \in H_B$ and $A' \in \bar{\mathcal{P}}(A)$,* $M_{A'}\phi_A F = \phi_{A'} F$,
3. $\dim H_B = d_{|B|}$, *where for $k \in \{2, \dots, n\}$, $d_k$ is the number of fixed-point free permutations (also called derangements) on a set with $k$ elements.*

*Example* 48. For $A = [\![4]\!]$, the multiresolution decomposition of $L(\mathfrak{S}_4)$ writes as

$$L(\mathfrak{S}_4) = \phi_{[\![4]\!]} \begin{pmatrix} H_{\{1,2,3,4\}} \\ H_{\{1,2,3\}} \oplus H_{\{1,2,4\}} \oplus H_{\{1,3,4\}} \oplus H_{\{2,3,4\}} \\ H_{\{1,2\}} \oplus H_{\{1,3\}} \oplus H_{\{1,4\}} \oplus H_{\{2,3\}} \oplus H_{\{2,4\}} \oplus H_{\{3,4\}} \\ H_{\emptyset} \end{pmatrix}.$$

The proof of Theorem 47 relies on two key properties, one about the spaces $H_B$ and the other about the operators $\phi_A$. We start with the latter, given by the following lemma. For $A \subset [\![n]\!]$ with $|A| = 1$ we set by convention $L(\Gamma(A)) = H_\emptyset$ and $M_A = M_\emptyset$.

**Lemma 49** (Commutation between marginal and wavelet synthesis operators)**.** *Let $A, B \in \mathcal{P}([\![n]\!])$, $F \in L(\Gamma(A))$ and $C \in \mathcal{P}([\![n]\!])$ such that $A \cup B \subset C$. Then $M_B \phi_C F = \phi_B M_{A\cap B} F$. In other words, the following diagram is commutative.*




*The diagram actually represents the restrictions of the operators to the involved spaces but we do not notify them for clarity's sake.*

Lemma 49 says in a way that the embedding operators $\phi_A$ commute with the marginal operators $M_B$. As its proof is purely technical, we leave it to the Appendix. We however provide an illustrating example.

*Example* 50. Let $A = \{1, 2, 3\}$, $B = \{1, 2, 4\}$ and $C = \{1, 2, 3, 4\}$. Then for $\pi = 123$ for instance,

$$M_B \phi_C \delta_\pi = M_{\{1,2,4\}} \phi_{\{1,2,3,4\}} \delta_{123} = \frac{1}{2} M_{\{1,2,4\}} \left[\delta_{4123} + \delta_{1234}\right] = \frac{1}{2}\left[\delta_{412} + \delta_{124}\right]$$

and $$\phi_B M_{A\cap B} \delta_\pi = \phi_{\{1,2,4\}} M_{\{1,2\}} \delta_{123} = \phi_{\{1,2,4\}} \delta_{12} = \frac{1}{2}\left[\delta_{412} + \delta_{124}\right].$$

Notice that with $A$, $B$, $C$ and $F$ as in Lemma 49, if $|A \cap B| \leq 1$ then $M_B \phi_C F = \phi_B M_\emptyset F$. Now by definition, $\phi_B M_\emptyset F$ is the constant function on $L(\Gamma(B))$ equal to $\sum_{\pi \in \Gamma(A)} F(\pi)$. This means that when $|A \cap B| \leq 1$, $M_B \phi_C F$ does not contain any information about $F$ besides its mean value. More generally, Lemma 49 implies the localization properties of the operators $\phi_A$ and therefore the MRA. A deeper interpretation is developed in Subsection 4.1.3. In practice, we use Lemma 49 to prove, for $A \in \bar{\mathcal{P}}([\![n]\!])$, the three following properties.

1. For $B \in \bar{\mathcal{P}}(A)$, $\phi_A$ is injective on $H_B$, *i.e.* $\ker \phi_A \cap H_B = \{0\}$.
2. For $B \in \bar{\mathcal{P}}(A)$, $F \in H_B$ and $A' \in \bar{\mathcal{P}}(A)$, $M_{A'}\phi_A F = \phi_{A'} F$.
3. The sum of spaces $(\phi_A(H_B))_{B \in \bar{\mathcal{P}}(A)}$ is direct.

*Proof.* We prove each property separately.

1. Let $F \in \ker \phi_A \cap H_B$. Applying Lemma 49 to $A, B := B$ and $C := A$ gives

$$\phi_B M_B F = M_B \phi_A F \qquad i.e. \qquad F = 0 \qquad \text{because } F \in \ker \phi_A,$$

   which concludes the proof.

2. Applying Lemma 49 to $A := B$, $B := A'$ and $C := A$ gives

$$M_{A'}\phi_A F = \phi_{A'} M_{B \cap A'} F.$$

   If $B \subset A'$ then $B \cap A' = B$ and one obtains $M_{A'}\phi_A F = \phi_{A'} M_B F = \phi_{A'} F$ because $F \in L(\Gamma(B))$. If $B \not\subset A'$ then $B \cap A' \subsetneq B$ and $M_{B \cap A'} F = 0$ because $F \in H_B$. Hence $M_{A'}\phi_A F = 0 = \phi_{A'} F$.

3. Let $(F_B)_{B \in \bar{\mathcal{P}}(A)} \in \bigoplus_{B \in \bar{\mathcal{P}}(A)} H_B$ such that

$$\sum_{B \in \bar{\mathcal{P}}(A)} \phi_A F_B = 0. \tag{4.1}$$

   We need to show that $F_B = 0$ for each $B \in \bar{\mathcal{P}}(A)$. We do it recursively on $|B|$ by applying property 2. to (4.1) for different subsets $A'$. First, applying $M_\emptyset$ cancels all the terms $\phi_A F_B$ for $B \in \mathcal{P}(A)$, leading to $F_\emptyset = 0$. Then for any $A' \subset A$ with $|A'| = 2$, applying $M_{A'}$ cancels all the terms $\phi_A F_B$ for $B \in \mathcal{P}(A) \setminus \{A'\}$, leading to $F_{A'} = 0$. The proof is concluded by induction.

□

The second key ingredient of the proof of Theorem 47 is the following theorem. We recall that for $k \in \{2, \dots, n\}$, $d_k$ is the number of derangements on a set of $k$ elements.

**Theorem 51** (Dimension of the space $H_{[\![k]\!]}$)**.** *For $k \in \{2, \dots, n\}$, $\dim H_{[\![k]\!]} = d_k$.*

Theorem 51 is proved in Reiner et al. (2014), where $H_{[\![k]\!]}$ is denoted by $\ker \pi_{[\![k]\!]}$ (see proposition 6.8 and corollary 6.15). As simple as it may seem, this result is far from being trivial. It is actually shown in Reiner et al. (2014) that $H_{[\![k]\!]}$ is isomorphic to the top homology space of the complex of injective words on $[\![k]\!]$. The calculation of the dimension of the latter relies on the Hopf trace formula for virtual characters and the topological properties of the partial order of subword inclusion, proved in several contributions of the algebraic topology literature (see Farmer, 1978; Björner and Wachs, 1983; Reiner and Webb, 2004).

*Example* 52. One recovers the values $\dim H_{\{1,2\}} = 1$ and $\dim H_{\{1,2,3\}} = 2$. Some more values of $d_k$ are given in Table 4.1.

Theorem 51 enables to conclude the proof of Theorem 47 with a dimensional argument. First observe that for $k \in \{2, \dots, n\}$, all the spaces $H_B$ for $B \subset [\![n]\!]$ with $|B| = k$ are isomorphic to

| $k$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| $k!$ | 1 | 1 | 2 | 6 | 24 | 120 | 720 |
| $d_k$ | 1 | 0 | 1 | 2 | 9 | 44 | 265 |

Table 4.1: Values of $k!$ and $d_k$

$H_{[\![k]\!]}$ (this is obvious, but also properly established by Proposition 61). Thus $\dim H_B = d_{|B|}$ for all $B \in \bar{\mathcal{P}}([\![n]\!])$. Combining this result with Properties 1 and 3, one obtains for any $A \in \mathcal{P}([\![n]\!])$,

$$|A|! = \dim L(\Gamma(A)) \geq \dim \bigoplus_{B \in \bar{\mathcal{P}}(A)} \phi_A(H_B) \geq \sum_{B \in \bar{\mathcal{P}}(A)} d_{|B|} = \sum_{k=0}^{|A|} \binom{|A|}{k} d_k = |A|!, \tag{4.2}$$

where the last equality is a classic result in elementary combinatorics. All the inequalities in (4.2) are therefore equalities, and the proof of Theorem 47 is finished.

### 4.1.3 Interpretation of the embedding operators $\phi_A$

Here we provide some more insights about the embedding operators $\phi_A$. Their definition is indeed not intuitive as mentioned previously. For $A, B \in \mathcal{P}([\![n]\!])$ with $B \subset A$, the most natural way to embed a Dirac function $\delta_\pi$ with $\pi \in \Gamma(B)$ into $L(\Gamma(A))$ would rather be to map it to the uniform distribution over all the rankings on $A$ that extend $\pi$, that is to use the following operator

$$\phi'_A \quad : \quad \delta_\pi \quad \mapsto \quad \frac{|B|!}{|A|!} \sum_{\sigma \in \pi,\ \pi \subset \sigma} \delta_\sigma. \tag{4.3}$$

*Example* 53. For $\pi = 42$ and $A = [\![4]\!]$:

$$\phi_A \delta_\pi = \frac{1}{6}\left[\delta_{1342} + \delta_{3142} + \delta_{1423} + \delta_{3421} + \delta_{4213} + \delta_{4231}\right]$$

$$\phi'_A \delta_\pi = \frac{1}{12}\left[\delta_{1342} + \delta_{3142} + \delta_{1423} + \delta_{3421} + \delta_{4213} + \delta_{4231} + \delta_{1432} + \delta_{3412} + \delta_{4132} + \delta_{4312} + \delta_{4123} + \delta_{4321}\right]$$

The operator $\phi'_A$ is used implicitly in shuffling interpretations of rankings (see Diaconis, 1988; Huang and Guestrin, 2012). It corresponds to mapping a ranking $\pi$ to the uniform distribution over all the possible shuffles between $\pi$ and a random ranking on $\Gamma(A \setminus B)$. Notice also that for $A = [\![n]\!]$, $\phi'_n : \pi \mapsto (|\pi|!/n!)\mathbb{1}_{\mathfrak{S}_n(\pi)}$. In other words, $\phi'_{[\![n]\!]}$ maps an incomplete ranking to the uniform distribution on the set of its linear extensions. It is thus also involved implicitly in the approaches introduced in Yu et al. (2002), Kondor and Barbosa (2010) and Sun et al. (2012) described in Subsection 3.1.5. For these two reasons, $\phi'_A$ can be considered as the most intuitive embedding operator. As a matter of fact, one can show that $\phi'_A$ also establishes an isomorphism between the spaces $\bigoplus_{B \in \bar{\mathcal{P}}(A)} H_B$ and $L(A)$, leading to another multiresolution decomposition (this is done in Section 6.2). This decomposition does not however satisfies the localization properties of Theorem 47. This is because the operator $\phi'_A$ does not satisfy the key Lemma 49, whereas the embedding operator $\phi_A$ does.

*Example* 54. Coming back to Example 50 with $A = \{1, 2, 3\}$, $B = \{1, 2, 4\}$ and $C = \{1, 2, 3, 4\}$, we recall that, for $\pi = 123$ for instance,

$$M_B \phi_C \delta_\pi = \frac{1}{2} M_{\{1,2,4\}}\left[\delta_{4123} + \delta_{1234}\right] = \frac{1}{2}\left[\delta_{412} + \delta_{124}\right],$$

$$\phi_B M_{A \cap B} \delta_\pi = \phi_{\{1,2,4\}} \delta_{12} = \frac{1}{2}\left[\delta_{412} + \delta_{124}\right].$$

By contrast,

$$M_B\phi'_C\delta_\pi = \frac{1}{4}M_{\{1,2,4\}}\left[\delta_{4123} + \delta_{1423} + \delta_{1243} + \delta_{1234}\right] = \frac{1}{4}\left[\delta_{412} + \delta_{142} + 2\delta_{124}\right]$$
$$\phi'_B M_{A\cap B}\delta_\pi = \phi_{\{1,2,4\}}\delta_{12} = \frac{1}{3}\left[\delta_{412} + \delta_{142} + \delta_{124}\right].$$

Even if the operators $\phi'_A$ were normalized differently, they would still not satisfy Lemma 49. In the example, the difference comes from the fact that the element $\delta_{1243}$ leads to an additional term $\delta_{124}$ in the end.

We now develop a more intuitive interpretation of the localization properties induced by the operators $\phi_A$. Let us consider for instance $\pi = 12 \in \Gamma(\{1,2\})$ and $C = [\![5]\!]$, and let $\sigma \in \Gamma([\![5]\!])$ be a ranking that extends $\pi$. It induces rankings on all subsets $B \in \mathcal{P}([\![5]\!])$ with in particular $\sigma_{|\{1,2\}} = \pi$. Now consider a perturbation that changes $\sigma$ to $\sigma'$ such that $\sigma'_{|\{1,2\}} = 21$. It necessarily changes the relative positions of elements 1 and 2 in $\sigma$ and more generally in all the subwords of $\sigma$ that contain 1 and 2. The question is then: how does it affect the other induced rankings $\sigma'_{|B}$ for $B \in \mathcal{P}([\![5]\!])$ such that $\{1,2\} \not\subset B$? If $B \cap \{1,2\} = \emptyset$, $\sigma'_{|B}$ is different from $\sigma_{|B}$ if and only if the perturbation also modifies the relative order of some elements in $B$. This is independent from the action on 1 and 2. Now, for $B \in \mathcal{P}([\![5]\!])$ such that $|B \cap \{1,2\}| = 1$, the key observation is that it depends on the elements that are placed *between* 1 and 2 in $\sigma$. For instance if $\sigma = 41523$, any perturbation that changes the relative positions of 1 and 2 will necessarily impact the relative position of at least 1 and 5 or 2 and 5. By contrast, if $\sigma = 45123$ for instance, swapping elements 1 and 2 will not have any impact on $\sigma_{|B}$ for all $B$ such that $|B \cap \{1,2\}| = 1$. Therefore among the rankings that extend 12, only the ones in which 1 and 2 are adjacent can be perturbed such that only the ranking induced on $\{1,2\}$ is affected and not the ones on the subsets $B$ with $|B \cap \{1,2\}| \leq 1$. A similar interpretation holds for subsets of elements of any size. Developing a general theory of perturbations for rankings would certainly be an interesting future research direction.

## 4.2 MRA representation

We now introduce the MRA representation, constructed from the multiresolution decomposition established in Section 4.1. The terminology used here and throughout the article is borrowed from wavelet theory. Though it can appear peculiar at first reading, the analogy is explained at length in Subsection 4.2.3.

### 4.2.1 Vocabulary and definitions

The MRA representation is constituted of four main objects: *signal space*, *feature space*, *wavelet transform* and *embedding operators*.

**Signal space.** The MRA representation applies to functions of incomplete rankings, which are seen as "signals" in order to borrow the language of standard MRA in wavelet theory for interpretation purpose (refer to Mallat, 2008). Any space $L(\Gamma(A))$ for $A \in \bar{\mathcal{P}}([\![n]\!])$ is seen as a *local* signal space and they are all embedded into the *global* signal space defined by

$$L(\bar{\Gamma}_n) = \bigoplus_{A\in\bar{\mathcal{P}}([\![n]\!])} L(\Gamma(A)).$$

Elements of the signal space are seen as collections of functions $F = (F_A)_{A\in\bar{\mathcal{P}}([\![n]\!])}$. The global support of an element $F$ is the set $\mathbf{supp}(F) = \{A \in \bar{\mathcal{P}}([\![n]\!]) \mid F_A \neq 0\}$, and we usually identify an element $F$ with the collection restricted to its global support $(F_A)_{A\in\mathbf{supp}(F)}$.

**Feature space.** The feature space is defined by

$$\mathbb{H}_n = \bigoplus_{B\in\bar{\mathcal{P}}([\![n]\!])} H_B.$$

Since $\dim H_B = d_{|B|}$ (the number of derangements on a set of $|B|$ elements, see Theorem 47), one has $\dim \mathbb{H}_n = \sum_{k=0}^{n} \binom{n}{k} d_k = n!$ by elementary combinatorics. Elements of the feature space are viewed as collections of vectors $\mathbf{X} = (X_B)_{B\in\bar{\mathcal{P}}([\![n]\!])}$. The global support of an element $\mathbf{X}$ is the set $\mathbf{supp}(\mathbf{X}) = \{B \in \bar{\mathcal{P}}([\![n]\!]) \mid X_B \neq 0\}$, and we usually identify an element $\mathbf{X}$ with the collection restricted to its global support $(X_B)_{B\in\mathbf{supp}(\mathbf{X})}$.

**Wavelet transform.** We first construct the wavelet transform implicitly from Theorem 47 (an explicit recursive construction is detailed in Subsection 4.2.6). For any $A \in \bar{\mathcal{P}}([\![n]\!])$ and $F \in L(\Gamma(A))$, the latter establishes the existence of a unique element $(\Psi_B^A F)_{B\in\bar{\mathcal{P}}(A)} \in \bigoplus_{B\in\bar{\mathcal{P}}(A)} H_B$ such that

$$F = \sum_{B\in\bar{\mathcal{P}}(A)} \phi_A \Psi_B^A F. \tag{4.4}$$

Property (4.4) defines for any $B \in \bar{\mathcal{P}}([\![n]\!])$ the linear operator $\Psi_B : L(\bar{\Gamma}_n) \to H_B$ on each subspace $L(\Gamma(A))$ for $A \in \bar{\mathcal{P}}([\![n]\!])$ as the mapping

$$\Psi_B : F \mapsto \Psi_B^A F \quad \text{if } B \subset A \text{ and } 0 \text{ otherwise.} \tag{4.5}$$

The operator $\Psi_B$ is called the *wavelet projector* on $H_B$ for $B \in \mathcal{P}([\![n]\!])$. The wavelet projections $\Psi_B F$ of a signal $F \in L(\bar{\Gamma}_n)$ are considered as its features. The wavelet transform is then defined as the collection of all the wavelet projectors. In other words, it maps a signal $F$ to all its features.

**Definition 55** (Wavelet transform)**.** The wavelet transform is the operator $\Psi : L(\bar{\Gamma}_n) \to \mathbb{H}_n$ constructed from the operators $\Psi_B$ defined in (4.5) as

$$\Psi : F \mapsto (\Psi_B F)_{B\in\bar{\mathcal{P}}([\![n]\!])}.$$

**Embedding operators** The embedding operators are the $\phi_A$'s defined for each $A \in \bar{\mathcal{P}}([\![n]\!])$ in Definition 44. In the context of the MRA representation, they can also be considered as synthesis operators: they reconstruct a signal in the local space $L(\Gamma(A))$ from its features. This is summarized by the following properties, direct consequences of Property (4.4):

$$\begin{aligned} & \phi_A \Psi(F) = F && \text{for any } F \in L(\Gamma(A)), \\ \text{and} \quad & \Psi\phi_{[\![n]\!]}(\mathbf{X}) = \mathbf{X} && \text{for any } \mathbf{X} = (X_B)_{B\in\bar{\mathcal{P}}([\![n]\!])} \in \mathbb{H}_n. \end{aligned}$$

*Example* 56. In this example we illustrate the objects of the MRA representation for $n = 4$.

| **Signal space** | | **Feature space** |
|---|---|---|
| $L(\Gamma(\{1,2,3,4\}))$ | | $H_{\{1,2,3,4\}}$ |
| $L(\Gamma(\{1,2,3\})) \oplus L(\Gamma(\{1,2,4\}))$ | **Wavelet transform** | $H_{\{1,2,3\}} \oplus H_{\{1,2,4\}}$ |
| $L(\Gamma(\{1,3,4\})) \oplus L(\Gamma(\{2,3,4\}))$ | $\Psi \longrightarrow$ | $H_{\{1,3,4\}} \oplus H_{\{2,3,4\}}$ |
| $L(\Gamma(\{1,2\})) \oplus L(\Gamma(\{1,3\})) \oplus L(\Gamma(\{1,4\}))$ | $\longleftarrow$ **Embedding operators** $(\phi_A)_{A\in\bar{\mathcal{P}}([\![4]\!])}$ | $H_{\{1,2\}} \oplus H_{\{1,3\}} \oplus H_{\{1,4\}}$ |
| $L(\Gamma(\{2,3\})) \oplus L(\Gamma(\{2,4\})) \oplus L(\Gamma(\{3,4\}))$ | | $H_{\{2,3\}} \oplus H_{\{2,4\}} \oplus H_{\{3,4\}}$ |
| $L(\Gamma(\bar{0}))$ | | $H_{\emptyset}$ |

By construction, the spaces $H_B$, the operators $\phi_A$ and the wavelet transform $\Psi$ satisfy, for all $A, B \in \bar{\mathcal{P}}([\![n]\!])$, $F_A \in L(\Gamma(A))$ and $X_B \in H_B$,

$$\begin{aligned} \Psi_B F_A &= 0 \qquad \text{if } B \not\subset A \\ \phi_A X_B &= 0 \qquad \text{if } B \not\subset A. \end{aligned}$$

For $B \in \bar{\mathcal{P}}([\![n]\!])$, we define the set $\mathcal{Q}(B) = \{A \in \bar{\mathcal{P}}([\![n]\!]) \mid B \subset A\}$. One then has, for any $A, B \in \mathcal{P}([\![n]\!])$, $F \in L(\bar{\Gamma}_n)$ and $\mathbf{X} \in \mathbb{H}_n$,

$$\Psi_B(F) = \sum_{A\in\mathcal{Q}(B)} \Psi_B F_A \qquad \text{and} \qquad \phi_A(\mathbf{X}) = \sum_{B\in\bar{\mathcal{P}}(A)} \phi_A X_B.$$

In words, this means that the wavelet projection of a signal on a space $H_B$ is the sum of the wavelet projections of the components of this signal on subsets that include $B$, and the embedding of a vector from the feature space in a local signal space $L(\Gamma(A))$ is the sum of the embeddings of all the components of the vector on subsets that are included in $A$.

### 4.2.2 Main properties

The strength of the MRA representation comes from the relationship between the wavelet transform, the embedding operator and the marginal operator, summarized in the following theorem. For any collection of subsets $\mathcal{S} \subset \bar{\mathcal{P}}([\![n]\!])$, we define the subspace of $\mathbb{H}_n$:

$$\mathbb{H}(\mathcal{S}) = \bigoplus_{B\in\mathcal{S}} H_B.$$

**Theorem 57** (Fundamental properties of the MRA representation)**.** *Let $A \in \bar{\mathcal{P}}([\![n]\!])$ and $F \in L(\Gamma(A))$. The MRA representation satisfies the following properties.*

- *$\Psi F$ is the unique element in $\mathbb{H}(\bar{\mathcal{P}}(A))$ such that*

$$F = \phi_A \Psi F = \sum_{B\in\bar{\mathcal{P}}(A)} \phi_A \Psi_B F. \tag{4.6}$$

- *For any $A' \in \bar{\mathcal{P}}(A)$,*

$$M_{A'} F = \phi_{A'} \Psi F \qquad \textit{or equivalently} \qquad \Psi_B M_{A'} F = \Psi_B F \textit{ for all } B \in \bar{\mathcal{P}}(A'). \tag{4.7}$$

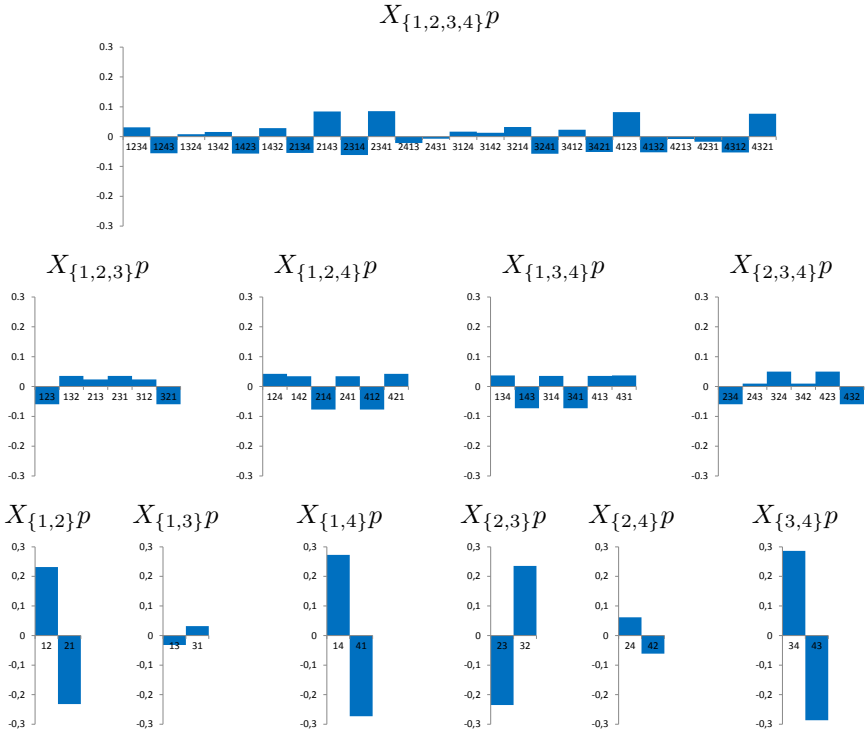



Figure 4.1: Wavelet projections of $p$ from the German dataset

*Proof.* Property (4.6) and the first part of Property (4.7) are direct consequences of Theorem 47 and Definition 55. To prove the second part of Property (4.7), observe that the first part applied to $F$ gives $M_{A'}F = \sum_{B\in\bar{\mathcal{P}}(A')} \phi_{A'}\Psi_B F$ and (4.6) applied to $M_{A'}F$ gives $M_{A'}F = \sum_{B\in\bar{\mathcal{P}}(A')} \phi_{A'}\Psi_B M_{A'}F$. The uniqueness of the decomposition concludes the proof. □

*Example* 58. We illustrate Theorem 57 on the distribution $p$ from the German dataset ($n = 4$). First, Figure 4.1 provides graphical representations for each wavelet projection $\Psi_B p$ for $B \in \mathcal{P}(\llbracket 4 \rrbracket)$ (we do not represent $\Psi_\emptyset p$ because it is simply equal to $\delta_{\bar{0}}$). Then Figure 4.2 illustrates the decomposition of $p$ with graphical representations for each component $\phi_{\llbracket 4 \rrbracket}\Psi_B p$ for $B \in \bar{\mathcal{P}}(\llbracket 4 \rrbracket)$, and of $M_{\{1,3,4\}}p$ with graphical representations for each component $\phi_{\{1,3,4\}}\Psi_B p$ for $B \in \bar{\mathcal{P}}(\{1,3,4\})$.

Theorem 57 has several implications in practice. First, Property (4.6) says that a function $F \in L(\Gamma(A))$ with $A \in \mathcal{P}(\llbracket n \rrbracket)$ can be reconstructed from its wavelet transform $\Psi F$. The latter thus contains all information related to $F$ or in other words, the knowledge of $\Psi F$ implies the knowledge of $F$. In addition, this piece of information is decomposed between all the wavelet projections $\Psi_B F$, and Property (4.7) says that this decomposition is consistent with the marginal operator: the marginal $M_{A'}F$ of $F$ on any subset $A' \in \bar{\mathcal{P}}(A)$ can be reconstructed from the wavelet transform of $F$ restricted to the subsets $B \in \bar{\mathcal{P}}(A')$. Figure 4.3 illustrates these properties for a ranking model $p$ over $\mathfrak{S}_3$ with marginals $P_{\{1,2\}}$, $P_{\{1,3\}}$ and $P_{\{2,3\}}$.

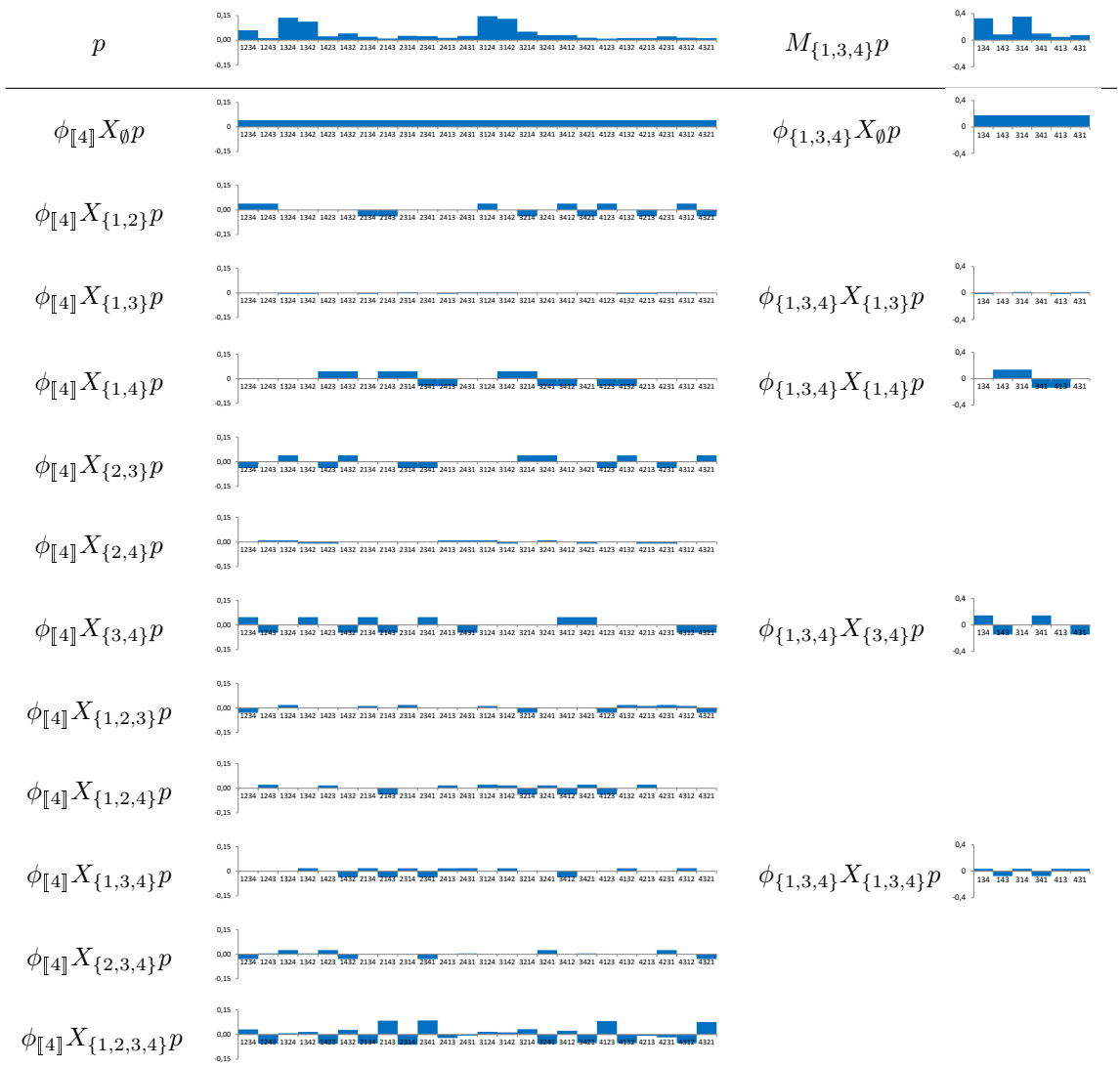



Figure 4.2: MRA decomposition of $p$ and $M_{\{1,3,4\}}p$ from the German dataset

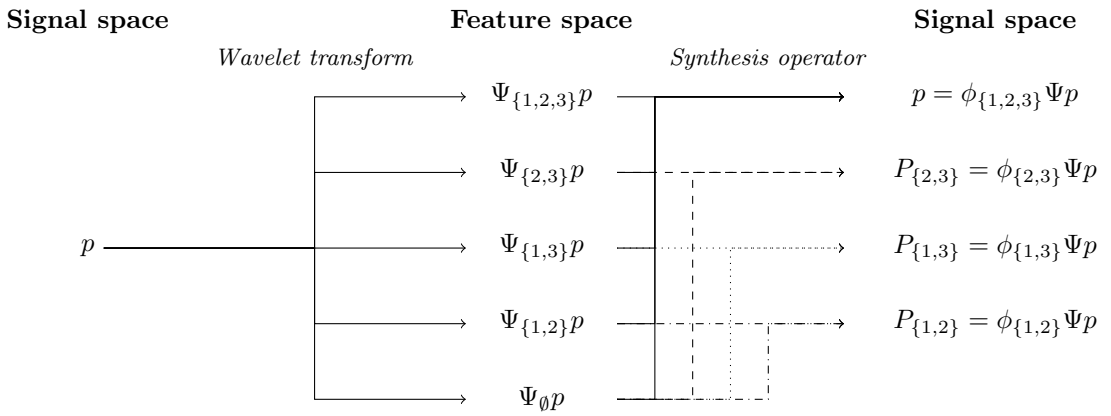



Figure 4.3: Illustration of Theorem 57 for $n = 3$

### 4.2.3 Multiresolution interpretation

We now show that the MRA representation exploits the natural multiscale structure of the marginals, justifying the use of terms "MRA representation" and "wavelet transform". Property 1. of Proposition 33 can be reformulated into the following relationships for any subsets $A, B \in \mathcal{P}(\llbracket n \rrbracket)$ with $B \subset A$:

$$(M_A)_{L(\Gamma(A))} = Id_{L(\Gamma(A))} \qquad \text{and} \qquad M_B M_A = M_B, \tag{4.8}$$

where $(M_A)_{L(\Gamma(A))}$ denotes the restriction of $M_A$ to $L(\Gamma(A))$ and $Id_{L(\Gamma(A))}$ is the identity operator on $L(\Gamma(A))$. Relationships (4.8) actually mean that the collection of linear spaces $(L(\Gamma(A)))_{A\in\mathcal{P}(\llbracket n \rrbracket)}$ together with the collection of linear operators $(M_A)_{A\in\mathcal{P}(\llbracket n \rrbracket)}$ form a *projective system*. The partial order associated with this projective system is the inclusion order on $\mathcal{P}(\llbracket n \rrbracket)$. It is canonically graded with the rank function $A \mapsto |A|$. This defines a notion of *scale* for the marginals, and this is why we call the projective system defined by relationships (4.8) the *multiscale structure* of the marginals. Figure 4.4 provides an illustration for $n = 4$.

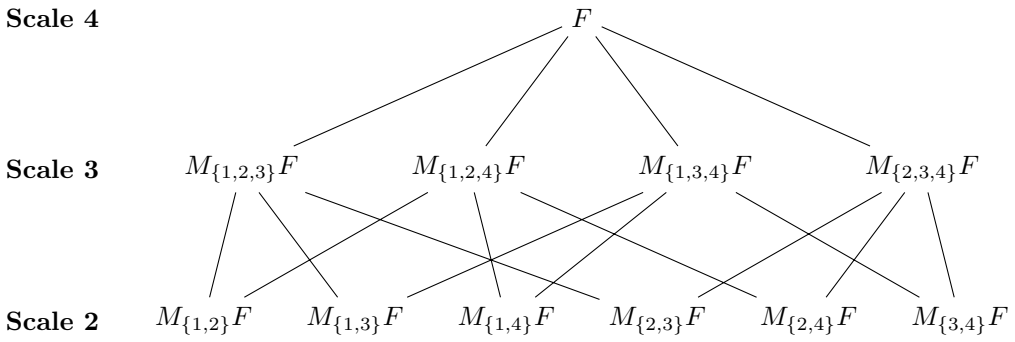



Figure 4.4: multiscale structure of the marginals of a function $F \in L(\Gamma(\llbracket 4 \rrbracket))$

From a practical point of view, the scale of a marginal corresponds to the number of items in the subset on which the marginal is considered. By equation (4.8), a marginal on a subset

$B \in \mathcal{P}(A)$ induces the marginals on all the subsets $C \in \mathcal{P}(B)$. The collection of marginals $(M_B F)_{B \subset A,\ |B|=k}$ for $F \in L(\Gamma(A))$ and $k \in \{2, \dots, |A|\}$ thus induces all the marginals on subsets $C \subset A$ with $|C| \leq k-1$. Hence we say that $(M_B F)_{B \subset A,\ |B|=k}$ contains all the information of $F$ at scale up to $k$. This notion of scale can be naturally compared to the usual notion in image analysis: its version in low resolution can be recovered from a higher resolution. The version of the image in the higher resolution thus contains more information than the version in low resolution.

The same as for images, the piece of information gained when increasing the scale corresponds to an additional level of details. For instance, if one has access to the triple-wise marginals of a ranking model $p$ then one has access to the information contained in the pairwise marginals plus the piece of information of scale 3. This decomposition can be further refined: marginals of the same scale on different subsets provide different additional pieces of information. For instance, compared to the marginal on $\{1,4\}$, the marginals on $\{1,2,4\}$ and $\{1,3,4\}$ both provide an additional but different level of details. Pursuing the analogy with images, this decomposition of the information into pieces related to subsets of items can be compared with the space decomposition of an image: for each resolution level, an image can be spatially decomposed into different components. Therefore, through their multiscale structure, the marginals of a function $F \in L(\Gamma(A))$ for $A \in \mathcal{P}(\llbracket n \rrbracket)$ each contain a part of its total information, both delimited in *scale* and in *space*.

The multiresolution representation enables to localize, in each of these parts, the component that is specific to the marginal. First, one has $\Psi_\emptyset F = (\sum_{\pi \in \Gamma(A)} F(\pi))\delta_{\bar{0}}$, this is proven in Subsection 4.2.6. The projection $\Psi_\emptyset F$ can thus be seen as containing the piece of information of $F$ at scale 0. Then for a pair $\{a,b\} \subset A$, applying Eq. (4.6) to $M_{\{a,b\}}F$ combined with (4.7) gives

$$\Psi_{\{a,b\}} F = M_{\{a,b\}} F - \phi_{\{a,b\}} \Psi_\emptyset F. \tag{4.9}$$

Hence, starting from $\Psi_\emptyset F$, $\Psi_{\{a,b\}}F$ contains the exact additional piece of information to recover $M_{\{a,b\}}F$. This is the part of information that is specific to $M_{\{a,b\}}F$. For a triple $\{a,b,c\} \subset A$, the same calculation gives

$$\Psi_{\{a,b,c\}} F = M_{\{a,b,c\}} F - \phi_{\{a,b,c\}} \left[ \Psi_\emptyset F + \Psi_{\{a,b\}} F + \Psi_{\{a,c\}} F + \Psi_{\{b,c\}} F \right]. \tag{4.10}$$

The projection $\Psi_{\{a,b,c\}}F$ of $F$ thus contains all the additional piece of information to get from the pairwise marginals $M_{\{a,b\}}F$, $M_{\{a,c\}}F$ and $M_{\{b,c\}}F$ to the triple-wise marginal $M_{\{a,b,c\}}F$. More generally, for $B \in \mathcal{P}(A)$, $\Psi_B F$ contains the piece of information that is specific to $M_B F$, or equivalently the part of the information of $F$ that is localized on scale $|B|$ on the subset $B$.

*Example* 59. Let $p$ be a ranking model over $\Gamma(\llbracket 3 \rrbracket)$ and $\Sigma$ a random permutation drawn from $p$. For clarity's sake, we denote by $\mathbb{P}[a_1 \succ \cdots \succ a_k]$ the probability of the event $\{\Sigma(a_1) < \cdots < \Sigma(a_k)\}$. One has for instance (see Subsection 4.2.6 for the general formulas):

$$\begin{aligned}
\mathbb{P}[2 \succ 1 \succ 3] &= p(213) \\
&= \phi_{\llbracket 3 \rrbracket} \Psi_\emptyset p(213) + \phi_{\llbracket 3 \rrbracket} \left[ \Psi_{\{1,2\}} p + \Psi_{\{1,3\}} p + \Psi_{\{2,3\}} p \right](213) + \phi_{\llbracket 3 \rrbracket} \Psi_{\llbracket 3 \rrbracket} p(213) \\
&= \tfrac{1}{6} + \tfrac{1}{2} \left[ \left( \mathbb{P}[2 \succ 1] - \tfrac{1}{2} \right) + \left( \mathbb{P}[1 \succ 3] - \tfrac{1}{2} \right) \right] + \Psi_{\llbracket 3 \rrbracket} p(213).
\end{aligned}$$

In this decomposition, the first term is the value of the uniform distribution over $\Gamma(\llbracket 3 \rrbracket)$, it represents information at scale 0. The second term represents the part of information at scale 2 of $p$ that is involved in the probability of the ranking $2 \succ 1 \succ 3$. The last term represents the part of information involved at scale 3.

In wavelet analysis on a Euclidean space, each wavelet coefficient of a function $f$ contains a specific part of information, localized in scale and space. In the present context, for $F \in L(\Gamma(A))$ with $A \in \mathcal{P}([\![n]\!])$ and $B \in \bar{\mathcal{P}}(A)$, the coefficient $\Psi_B F$ contains the part of information that is specific to the marginal $M_B F$, or in other words localized at scale $|B|$ and subset $B$. This is why we call the operator $\Psi$ the *wavelet transform* and more generally the construction the *MRA representation*.

This analogy is based on the concept of information localization, in space and scale. Traditional multiresolution analysis on $\mathbb{R}^d$ is also characterized by its interplay with translation and dilation operators: translations enable to "move" in one resolution level and dilations to change between resolution levels. To develop the analogy, we define the following spaces.

**Definition 60** (Space $H^k$)**.** For $k \in \{0, \dots, n\} \setminus \{1\}$, we define

$$H^k = \bigoplus_{B \subset [\![n]\!],\ |B|=k} H_B, \qquad \text{so that} \qquad \mathbb{H}_n = \bigoplus_{\substack{k=0 \\ k \neq 1}}^{n} H^k.$$

For $k \in \{0, \dots, n\} \setminus \{1\}$ the space $H^k$ localizes all relative rank information of scale $k$. Analogously to classic multiresolution analysis, one can "move" inside a space $H^k$ between different subsets of elements with the translation operators.

**Proposition 61.** *Let $A, B \in \bar{\mathcal{P}}([\![n]\!])$ and $\tau \in \mathfrak{S}_n$. The three following properties hold.*

1. $T_\tau \phi_A = \phi_{\tau(A)} T_\tau$
2. $T_\tau \Psi_B = \Psi_{\tau(B)} T_\tau$
3. $T_\tau(H_B) = H_{\tau(B)}$.

Beyond the insights it provides, Proposition 61 is very useful for computations (see Section 4.3). Refer to the Appendix for its proof. Notice that it directly implies that the spaces $H^k$ are invariant under translations: for $k \in \{0, \dots, n\} \setminus \{1\}$ and $\tau \in \mathfrak{S}_n$ one has

$$T_\tau(H^k) = T_\tau \left( \bigoplus_{B \subset [\![n]\!],\ |B|=k} H_B \right) = \bigoplus_{B \subset [\![n]\!],\ |B|=k} T_\tau(H_B) = \bigoplus_{B \subset [\![n]\!],\ |B|=k} H_{\tau(B)} = H^k. \tag{4.11}$$

This means in particular that $H^k$ is a representation of $\mathfrak{S}_n$ and admits a Fourier decomposition. This property is developed in details in Section 6.1. The relationship between the MRA and Fourier analysis is another common point with classic multiresolution analysis. By contrast, operators that enable to "move" between the scales like the dilation operators in classic multiresolution analysis do not exist in the present context.

*Remark* 62 (Non orthogonality of the MRA decomposition)*.* We point out that the MRA decomposition is not orthogonal. More specifically, for any $A \in \mathcal{P}([\![n]\!])$ the subspaces $(\phi_A(H_B))_{B \in \mathcal{P}(A)}$ of $L(\Gamma(A))$ are not two-by-two orthogonal. This is not the case either for the spaces $\phi_A(H^k) = \bigoplus_{B \subset A, |B|=k} \phi_A(H_B)$ for $k = 2, \dots, |A|$. Only the space $\phi_A(H_\emptyset) = \phi_A(H^0)$ is orthogonal to all the $\phi_A(H_B)$'s for $B \in \mathcal{P}(A)$. As a consequence, one has for $F \in L(\Gamma(A))$

$$\|F\|_A^2 = \|\phi_A \Psi_\emptyset F\|_A^2 + \left\| \sum_{B \in \mathcal{P}(A)} \phi_A \Psi_B F \right\|_A^2 \text{ with in general } \left\| \sum_{B \in \mathcal{P}(A)} \phi_A \Psi_B F \right\|_A^2 \neq \sum_{B \in \mathcal{P}(A)} \|\Psi_B F\|_B^2,$$

where $\|\cdot\|_B$ is the abbreviated notation for the Euclidean norm $\|\cdot\|_{\Gamma(B)}$ on $L(\Gamma(B))$ for any $B \in \mathcal{P}([\![n]\!])$ that we use here and throughout the rest of the thesis.

### 4.2.4 Approximation in the MRA representation

Traditional wavelet analysis is naturally used together with linear or nonlinear approximation (see for instance Donoho et al., 1996; DeVore, 1998; Mallat, 2008). To illustrate the general principles, let $\mathcal{B} = \{\phi_0\} \cup \{\psi_{j,k}\}_{1 \leq k \leq 2^{j-1}, 1 \leq j \leq m}$ be a wavelet basis of $\mathbb{R}^{2^m}$ with $m \geq 1$. Since $\mathcal{B}$ is orthonormal, any function $f \in \mathbb{R}^{2^m}$ satisfies

$$f = \langle f, \phi_0 \rangle \, \phi_0 + \sum_{j=1}^{m} \sum_{k=1}^{2^{j-1}} \langle f, \psi_{j,k} \rangle \, \psi_{j,k} \qquad \text{and} \qquad \|f\|^2 = \langle f, \phi_0 \rangle^2 + \sum_{j=1}^{m} \sum_{k=1}^{2^{j-1}} \langle f, \psi_{j,k} \rangle^2 ,$$

where $\| \cdot \|$ denotes here the canonical Euclidean norm on $\mathbb{R}^{2^m}$. The two usual approximation schemes are defined as follows.

- **Linear approximation.** For $J \in \{0, \ldots, m\}$, the linear approximation of $f$ at scale $J$ is

$$f_J = \langle f, \phi_0 \rangle \, \phi_0 + \sum_{j=1}^{J} \sum_{k=1}^{2^{j-1}} \langle f, \psi_{j,k} \rangle \, \psi_{j,k} \qquad \text{with error} \qquad \|f - f_J\|^2 = \sum_{j=J+1}^{m} \sum_{k=1}^{2^{j-1}} \langle f, \psi_{j,k} \rangle^2 .$$

- **Nonlinear approximation.** To define the nonlinear approximation, we first give an arbitrary labeling to the wavelet basis: $\mathcal{B} = \{\psi_i\}_{1 \leq i \leq 2^m}$. For $M \in \{1, \ldots, 2^m\}$, let $f_M$ be the linear combination of the form $\sum_{i \in I} \langle f, \psi_i \rangle \, \psi_i$ with $|I| = M$ with minimal error:

$$f_M = \sum_{i \in I(M)} \langle f, \psi_i \rangle \, \psi_i \qquad \text{with} \qquad I(M) = \underset{I \subset \{1, \ldots, 2^m\}, \; |I| = M}{\operatorname{argmin}} \left\| f - \sum_{i \in I} \langle f, \psi_i \rangle \, \psi_i \right\|^2 .$$

  The function $f_M$ is called the *best M-term approximation* of $f$. Because $\mathcal{B}$ is orthonormal, $f_M$ is obtained by keeping in $f$ only the terms with highest absolute value of their coefficient:

$$f_M = \sum_{l=1}^{M} f_{\mathcal{B}}[l] \psi[l] \qquad \text{and its error is} \qquad \|f - f_M\|^2 = \sum_{l=M+1}^{2^m} f_{\mathcal{B}}[l]^2,$$

  where $f_{\mathcal{B}}[l]$ denotes the coefficient $\langle f, \psi_i \rangle$ with $l^{th}$ highest absolute value and $\psi[l]$ denotes the associated basis element.

A function $f \in \mathbb{R}^{2^m}$ is well approximated by the linear approximation scheme if $\|f - f_J\|^2$ decreases rapidly with $J$ and by the nonlinear approximation scheme if $\|f - f_M\|^2$ decreases rapidly with $M$. Of course, functions that are well approximated by the linear scheme are also well approximated by the nonlinear scheme. The latter is thus more powerful in the sense that it provides good approximations for more functions. Its utility on real data depends however on the basis. Nonlinear approximation in a Fourier basis is rarely used, because the functions that are well approximated by it but not by linear approximation rarely appear in real data. In other words, real data that would be well approximated by the nonlinear scheme in a Fourier basis are usually also well approximated by the linear scheme. One strength of wavelet bases is precisely that many of the functions from real data that are not well approximated by the nonlinear scheme in a Fourier basis are well approximated in them (see for instance Donoho, 1993).

It is natural to study analogous approximation schemes for the MRA representation. This is what we do next, pointing out at the same time several important differences with classic

wavelet analysis. Let $F \in L(\Gamma(A))$ with $A \in \mathcal{P}([\![n]\!])$ be a function to approximate. Its wavelet decomposition writes as

$$F = \sum_{B \in \bar{\mathcal{P}}(A)} \phi_A \Psi_B F.$$

As running example, we take $p$ from the German dataset.

**Error measure.** As the MRA decomposition is not orthogonal, measuring the approximation error with the Euclidean norm is not necessarily the most natural. We consider more generally $l^r$ norms for $r \geq 1$, denoting by $\|\cdot\|_{A,r}$ the $l^r$ norm on $L(\Gamma(A))$ in this subsection. More importantly, a first difference with classic wavelet analysis is that depending on the application, $F$ can be seen as a signal itself, or as a function that generates signals in local spaces $L(\Gamma(B))$ for $B \in \bar{\mathcal{P}}(A)$ through its marginals. While in the first case one would naturally define the error of an approximation $\tilde{F} \in L(\Gamma(A))$ by $\|\tilde{F} - F\|_{A,r}$, an error for the second case would rather be of the form $\sum_{B \in \bar{\mathcal{P}}(A)} w_B \|M_B(\tilde{F} - F)\|_{B,r}$, where the $w_B$'s are weighting coefficient. Such an error naturally arises for instance in the statistical problem of estimating the marginals of a ranking model (see Section 5.2).

**Linear approximation.** With the notion of scale defined in Subsection 4.2.3 and by analogy with classic wavelet analysis, it is natural to consider the linear approximation scheme that keeps only the lower scales. We thus define the linear approximation at scale $k$ of $F$ by

$$F_k = \sum_{B \in \mathcal{P}(A), |B| \leq k} \phi_A \Psi_B F = \phi_A \Psi_\emptyset F + \sum_{j=2}^{k} \sum_{B \subset A, |B|=j} \phi_A \Psi_B F$$

for $k = 1, \ldots, n$. Table 4.2 shows the results of the linear approximation scheme on $p$ from the German dataset. Errors $l^2$ and $l^1$ correspond to classic errors measured by $\|\cdot\|_{[\![4]\!],2}$ and $\|\cdot\|_{[\![4]\!],1}$ respectively. Errors $l^2_m$ and $l^1_m$ are the average of errors on the marginals defined by $(1/11)\sum_{B \in \mathcal{P}([\![4]\!])} \|M_B(F_k - F)\|_{B,r}$ for $r = 1, 2$. What is certainly most interesting from Table 4.2 is that $p_3$ is a bad approximation of $p$ for the $l^2$ and $l^1$ errors compared to $p_2$ and even to $p_1$. It is however better for the $l^2_m$ and $l^1_m$ errors. This highlights the fact that the MRA representation does not provide the localization properties one can be used to. The fact that the components $\phi_A \Psi_B F$ do not contain information about $F$ "local in the space $\Gamma(A)$" can also be observed on Figure 4.2: their support is diffuse and not restricted to a few rankings only. The difference between the localization properties of the MRA and the ones of classic wavelet analysis is further developed in Section 4.3.

**Nonlinear approximation.** For $M = 1, \ldots, 2^n - n$ and a given error $\mathcal{E}$, the best $M$-term approximation of F is defined by

$$F_M = \sum_{B \in \mathcal{S}(M)} \phi_A \Psi_B F \qquad \text{with} \qquad \mathcal{S}(M) = \operatorname*{argmin}_{\mathcal{S} \subset \bar{\mathcal{P}}(A),\ |\mathcal{S}|=M} \mathcal{E}\left(F - \sum_{B \in \mathcal{S}} \phi_A \Psi_B F\right).$$

As the MRA representation is not orthogonal, finding $F_M$ is not easy, even if $\mathcal{E} = \|\cdot\|_{[\![n]\!],2}$. Indeed even in this case, keeping the $M$ terms in $F$ with highest value of $\|\Psi_B F\|_{B,2}$ does not give $F_M$ in general. We however use here the latter procedure because it enables us to highlight another specificity of the present setting: the wavelet projections are vectors, not scalars. The question is then: on what quantity should we base the selection of the components? As a heuristic, we propose to consider for a subset $B \in \mathcal{P}(A)$ the quantity $\mathcal{E}(\phi_A \Psi_B F)$. For $\mathcal{E} = \|\cdot\|_{A,r}$ and $\mathcal{E} = \sum_{A' \in \mathcal{P}(A)} w_{A'} \|M_{A'} \cdot\|_{A',r}$, this quantity is computed using the following Lemma.

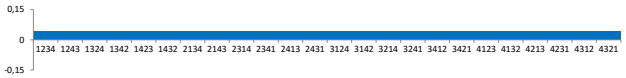
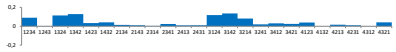
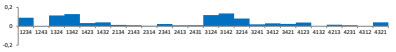
| **Scale** | $l^2$ | $l^2_m$ | $l^1$ | $l^1_m$ | $p_k$ |
|---|---|---|---|---|---|
| 1 | 0.21 | 0.27 | 0.77 | 0.51 |  |
| 2 | 0.15 | 0.06 | 0.66 | 0.16 |  |
| 3 | 0.24 | 0.02 | 0.99 | 0.09 |  |
| 4 | 0 | 0 | 0 | 0 |  |

Table 4.2: Linear approximation on the German dataset

| $M$ | $p_M$ |
|---|---|
| 3 |  |
| 6 |  |
| 9 |  |
| 12 |  |

(a) $l^2$ error

| $M$ | $p_M$ |
|---|---|
| 3 |  |
| 6 |  |
| 9 |  |
| 12 |  |

(b) $l^1$ error

| $M$ | $p_M$ |
|---|---|
| 3 |  |
| 6 |  |
| 9 |  |
| 12 |  |

(c) $l^2_m$ error

| $M$ | $p_M$ |
|---|---|
| 3 |  |
| 6 |  |
| 9 |  |
| 12 |  |

(d) $l^1_m$ error

Figure 4.5: Nonlinear approximation on the German dataset

| Rank | $l^2$ | $l^1$ | $l^2_m$ | $l^1_m$ | | Borda Count |
|---|---|---|---|---|---|---|
| 1 | $\{1,2,3,4\}$ | $\emptyset$ | $\emptyset$ | $\emptyset$ | | $\emptyset$ |
| 2 | $\emptyset$ | $\{1,2,3,4\}$ | $\{3,4\}$ | $\{3,4\}$ | | $\{3,4\}$ |
| 3 | $\{3,4\}$ | $\{3,4\}$ | $\{1,4\}$ | $\{1,4\}$ | | $\{1,4\}$ |
| 4 | $\{1,4\}$ | $\{1,4\}$ | $\{2,3\}$ | $\{2,3\}$ | | $\{1,2,3,4\}$ |
| 5 | $\{2,3\}$ | $\{2,3\}$ | $\{1,2\}$ | $\{1,2\}$ | | $\{2,3\}$ |
| 6 | $\{1,2\}$ | $\{1,2\}$ | $\{2,4\}$ | $\{1,2,3,4\}$ | $\longrightarrow$ | $\{1,2\}$ |
| 7 | $\{1,2,4\}$ | $\{1,2,4\}$ | $\{1,2,3,4\}$ | $\{1,2,4\}$ | | $\{1,2,4\}$ |
| 8 | $\{1,3,4\}$ | $\{1,3,4\}$ | $\{1,2,4\}$ | $\{1,3,4\}$ | | $\{1,3,4\}$ |
| 9 | $\{2,3,4\}$ | $\{2,3,4\}$ | $\{1,3,4\}$ | $\{2,4\}$ | | $\{2,4\}$ |
| 10 | $\{1,2,3\}$ | $\{1,2,3\}$ | $\{2,3,4\}$ | $\{2,3,4\}$ | | $\{2,3,4\}$ |
| 11 | $\{2,4\}$ | $\{2,4\}$ | $\{1,2,3\}$ | $\{1,2,3\}$ | | $\{1,2,3\}$ |
| 12 | $\{1,3\}$ | $\{1,3\}$ | $\{1,3\}$ | $\{1,3\}$ | | $\{1,3\}$ |

Figure 4.6: Orderings of the components for different errors on the German dataset

**Lemma 63.** *Let $A \in \bar{\mathcal{P}}([\![n]\!])$, $(X_B)_{B\in\bar{\mathcal{P}}(A)} \in \mathbb{H}(A)$ and $r \geq 1$.*

1. *For all $B \in \mathcal{P}(A)$,*

$$\|\phi_A X_\emptyset\|_{A,r} = \frac{1}{|A|!^{1-1/r}}|X_\emptyset(\bar{0})| \qquad \text{and} \qquad \|\phi_A X_B\|_{A,r} = \frac{1}{(|A|-|B|+1)!^{1-1/r}}\|X_B\|_{B,r}.$$

2. *Let $(w_{A'})_{A'\in\mathcal{P}(A)} \in \mathbb{R}^{2^{|A|}-|A|-1}$. For all $B \in \mathcal{P}(A)$,*

$$\sum_{A'\in\mathcal{P}(A)} w_{A'}\|M_{A'}\phi_A X_\emptyset\|_{A',r} = \left(\sum_{A'\in\mathcal{P}(A)} \frac{w_{A'}}{|A'|!^{1-1/r}}\right)|X_\emptyset(\bar{0})|$$

$$\text{and} \qquad \sum_{A'\in\mathcal{P}(A)} w_{A'}\|M_{A'}\phi_A X_B\|_{A',r} = \left(\sum_{A'\in\mathcal{P}(A)\cap\mathcal{Q}(B)} \frac{w_{A'}}{(|A'|-|B|+1)!^{1-1/r}}\right)\|X_B\|_{B,r}$$

*where we recall that $\mathcal{Q}(B) = \{A \in \mathcal{P}([\![n]\!]) \mid B \subset A\}$ for any $B \in \mathcal{P}([\![n]\!])$.*

Refer to the Appendix for the proof of Lemma 63. Figure 4.5 shows some of the nonlinear approximations of $p$ from the German dataset for several errors $\mathcal{E}$. An interesting question that usually comes with nonlinear approximation is: can we "summarize" $F$ into a small number of its components (see Diaconis, 1989). Unfortunately this question is ill-posed in the present setting because the MRA representation is not orthogonal and there are several ways of measuring how a sum of some components would approximate it or equivalently "summarize" it. Figure 4.6 shows the ordering of the components with respect to the value $\mathcal{E}(\phi_{[\![4]\!]}\Psi_B p)$ for several errors $\mathcal{E}$, as well as the aggregated ordering obtained by Borda Count (refer to Subsection 6.3.3 for the definition), which can be seen as providing the "average representativeness" of a component.

*Remark* 64 (Algorithms for nonlinear approximation). One could certainly apply methods from nonlinear approximation in a general dictionary (such as Orthogonal Matching Pursuit) but they should be adapted in the present setting where there is no dictionary *per se*. Otherwise, one could apply methods from combinatorial optimization such as *branch and bound* algorithms. This could constitute an interesting direction for future work.

### 4.2.5 Solving linear systems involving the marginal operator

One of the main consequences of Theorem 57 is that the MRA representation "simultaneously block-diagonalizes" the marginal operators $M_A$. To be more specific, let $\mathcal{M}_A : \mathbb{H}_n \to \mathbb{H}_n$ be the operator defined by $\mathcal{M}_A = \Psi M_A \phi_{[\![n]\!]}$ for $A \in \mathcal{P}([\![n]\!])$. The following proposition is a direct consequence of Theorem 57.

**Proposition 65** (Marginal operator in the feature space)**.** *Let $A \in \mathcal{P}([\![n]\!])$. For all $F \in L(\bar{\Gamma}_n)$,*

$$\Psi M_A F = \mathcal{M}_A \Psi F.$$

*In other words, the operator $\mathcal{M}_A$ is such that the following diagram is commutative.*

$$\begin{array}{ccc} L(\bar{\Gamma}_n) & \xrightarrow{\Psi} & \mathbb{H}_n \\ {\scriptstyle M_A}\downarrow & & \downarrow{\scriptstyle \mathcal{M}_A} \\ L(\Gamma(A)) & \xrightarrow[\Psi]{} & \mathbb{H}_n \end{array}$$

*Proof.* Let $A \in \mathcal{P}([\![n]\!])$ and $F \in L(\bar{\Gamma}_n)$. By definition of the operator $\mathcal{M}_A$,

$$\mathcal{M}_A \Psi F = \Psi M_A \phi_{[\![n]\!]} \Psi F.$$

Now, applying Property (4.7) successively to $\phi_{[\![n]\!]} \Psi F$ and $F$ gives

$$M_A \phi_{[\![n]\!]} \Psi F = \phi_A \Psi \phi_{[\![n]\!]} \Psi F = \phi_A \Psi F = M_A F,$$

where we recall that $\Psi \phi_{[\![n]\!]} \mathbf{X} = \mathbf{X}$ for any $\mathbf{X} \in \mathbb{H}_n$. Hence $\mathcal{M}_A \Psi F = \Psi M_A F$. □

Proposition 65 says that applying the operator $\mathcal{M}_A$ in the feature space is equivalent to applying the marginal operator $M_A$ in the signal space. This is why we call $\mathcal{M}_A$ the *marginal operator in the feature space*. Now, Theorem 57 also implies that this operator is actually a simple projection.

**Proposition 66** (Simultaneous block-diagonalization)**.** *For $A \in \mathcal{P}([\![n]\!])$, $\mathcal{M}_A$ is the projection on $\mathbb{H}(\bar{\mathcal{P}}(A))$: for any $(X_B)_{B \in \bar{\mathcal{P}}([\![n]\!])} \in \mathbb{H}_n$,*

$$\mathcal{M}_A \left( (X_B)_{B \in \bar{\mathcal{P}}([\![n]\!])} \right) = (X_B)_{B \in \bar{\mathcal{P}}(A)}.$$

*Equivalently, the matrix of $\mathcal{M}_A$ in any basis of $\mathbb{H}_n$ consistent with the decomposition $\bigoplus_{B \in \bar{\mathcal{P}}([\![n]\!])} H_B$ is of the form*

$$\begin{array}{c} \\ H_\emptyset \\ \vdots \\ H_{[\![n]\!]} \end{array} \begin{array}{c} \begin{array}{ccc} H_\emptyset & \cdots & H_{[\![n]\!]} \end{array} \\ \begin{bmatrix} \mathbf{m}_\emptyset & \cdots & 0 \\ \vdots & \ddots & 0 \\ 0 & \cdots & \mathbf{m}_{[\![n]\!]} \end{bmatrix} \end{array},$$

*where for $B \in \bar{\mathcal{P}}([\![n]\!])$, $\mathbf{m}_B = \mathbf{I}_B$, the matrix of the identity operator $Id_{H_B}$ on $H_B$, if $B \subset A$ and $\mathbf{m}_B = 0$ otherwise.*

*Proof.* Let $A \in \mathcal{P}([\![n]\!])$ and $\mathbf{X} \in \mathbb{H}_n$. Applying Property (4.7) to $\phi_{[\![n]\!]}\mathbf{X}$ one obtains

$$\mathcal{M}_A(\mathbf{X}) = \Psi M_A \phi_{[\![n]\!]}(\mathbf{X}) = \Psi \phi_A(\mathbf{X}) = \Psi \sum_{B \in \bar{\mathcal{P}}(A)} \phi_A X_B = (X_B)_{B \in \bar{\mathcal{P}}(A)},$$

which concludes the proof. □

*Example* 67. The matrix of $\mathcal{M}_{\{1,2,4\}}$ in any basis of $\mathbb{H}_4$ consistent with the decomposition $\bigoplus_{B \in \bar{\mathcal{P}}([\![4]\!])} H_B$ is equal to

| | $H_\emptyset$ | $H_{\{1,2\}}$ | $H_{\{1,3\}}$ | $H_{\{1,4\}}$ | $H_{\{2,3\}}$ | $H_{\{2,4\}}$ | $H_{\{3,4\}}$ | $H_{\{1,2,3\}}$ | $H_{\{1,2,4\}}$ | $H_{\{1,3,4\}}$ | $H_{\{2,3,4\}}$ | $H_{[\![4]\!]}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $H_\emptyset$ | $\mathbf{I}_\emptyset$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $H_{\{1,2\}}$ | 0 | $\mathbf{I}_{\{1,2\}}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $H_{\{1,3\}}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $H_{\{1,4\}}$ | 0 | 0 | 0 | $\mathbf{I}_{\{1,4\}}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $H_{\{2,3\}}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $H_{\{2,4\}}$ | 0 | 0 | 0 | 0 | 0 | $\mathbf{I}_{\{2,4\}}$ | 0 | 0 | 0 | 0 | 0 | 0 |
| $H_{\{3,4\}}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $H_{\{1,2,3\}}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $H_{\{1,2,4\}}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | $\mathbf{I}_{\{1,2,4\}}$ | 0 | 0 | 0 |
| $H_{\{1,3,4\}}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $H_{\{2,3,4\}}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $H_{[\![4]\!]}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Proposition 66 says at the same time that the marginal operator in the MRA representation boils down to a simple filter, and that all the marginal operators are "block-diagonalized" in the MRA representation. These properties mean that the MRA representation is best fitted to solve linear systems that involve the marginal operator. This is formalized in the following theorem. For any collection $\mathcal{S} \subset \mathcal{P}([\![n]\!])$, we set

$$\bar{\mathcal{P}}(\mathcal{S}) := \bigcup_{A \in \mathcal{S}} \bar{\mathcal{P}}(A).$$

**Theorem 68** (Solutions to linear systems)**.** *Let $A \in \mathcal{P}([\![n]\!])$ and $F_0 \in L(\Gamma(A))$.*

- *For $A' \in \mathcal{P}(A)$, the solutions to the problem*

$$\text{Find } F \in L(\Gamma(A)) \text{ such that } M_{A'}F = M_{A'}F_0 \tag{4.12}$$

*are all of the form*

$$\phi_A \sum_{B \in \bar{\mathcal{P}}(A')} \Psi_B F_0 \quad + \quad \phi_A \mathbf{X},$$

*with $\mathbf{X} \in \mathbb{H}(\bar{\mathcal{P}}(A) \setminus \mathcal{P}(A'))$. In particular the space of solutions has dimension $\dim \mathbb{H}(\bar{\mathcal{P}}(A) \setminus \mathcal{P}(A')) = |A|! - |A'|!$.*

- *More generally for $\mathcal{S} \subset \mathcal{P}(A)$, the solutions to the problem*

$$\text{Find } F \in L(\Gamma(A)) \text{ such that } M_{A'}F = M_{A'}F_0 \text{ for all } A' \in \mathcal{S} \tag{4.13}$$

*are all of the form*

$$\phi_A \sum_{B \in \bar{\mathcal{P}}(\mathcal{S})} \Psi_B F_0 \quad + \quad \phi_A \mathbf{X}$$

*with $\mathbf{X} \in \mathbb{H}(\bar{\mathcal{P}}(A) \setminus \mathcal{P}(\mathcal{S}))$. In particular the space of solutions has dimension $\dim \mathbb{H}(\bar{\mathcal{P}}(A) \setminus \mathcal{P}(\mathcal{S})) = |A|! - \sum_{A' \in \mathcal{S}} d_{|A'|}$.*

*Proof.* It is sufficient to prove the theorem for problem (4.13). Let $F \in L(\Gamma(A))$. For any $A' \in \mathcal{S}$,

$$\begin{array}{rcll} M_{A'}F = M_{A'}F_0 & \Leftrightarrow & \Psi M_{A'}F = \Psi M_{A'}F_0 & \text{by Theorem 57} \\ & \Leftrightarrow & \mathcal{M}_{A'}\Psi F = \mathcal{M}_{A'}\Psi F_0 & \text{by Proposition 65} \\ & \Leftrightarrow & \Psi_B F = \Psi_B F_0 \text{ for all } B \in \bar{\mathcal{P}}(A') & \text{by Proposition 66.} \end{array}$$

Thus $M_{A'}F = M_{A'}F_0$ for all $A' \in \mathcal{S}$ if and only if $\Psi_B F = \Psi_B F_0$ for all $B \in \bar{\mathcal{P}}(\mathcal{S})$. Applying Theorem 57 one last time concludes the proof. □

*Example* 69. We illustrate Theorem 68 for the ranking model $p$ from the German dataset. Let us assume that one does not know the ranking model $p$, but knows exactly some of its marginals $M_A p$ for subsets $A$ in the observation design $\mathcal{A} = \{\{1,3\},\{2,4\},\{3,4\},\{1,2,3\},\{1,3,4\}\}$. One then has $\mathcal{P}(\llbracket 4 \rrbracket) \setminus \bar{\mathcal{P}}(\mathcal{A}) = \{\{1,2,4\},\{2,3,4\},\{1,2,3,4\}\}$. Theorem 68 thus tells us that the functions on $\mathfrak{S}_4$ that have the same marginal as $p$ for all subsets $A \in \mathcal{A}$ are of the form

$$F = \phi_{\llbracket 4 \rrbracket} \sum_{B \in \bar{\mathcal{P}}(\mathcal{A})} \Psi_B p + \phi_{\llbracket 4 \rrbracket}\left[X_{\{1,2,4\}} + X_{\{2,3,4\}} + X_{\{1,2,3,4\}}\right] \qquad \text{with } X_B \in H_B,$$

where the $X_B$'s can be arbitrary. The set composed of such functions is therefore a linear space of dimension $d_4 + 2d_3 = 13$. Examples of such functions with their marginals are represented in Figure 4.7. The graphs on the left represent the function with $X_B = 0$ and the graphs on the right represent a function obtained with $X_B$'s sampled randomly.

### 4.2.6 Explicit construction of the wavelet transform

Definition 55 relies on an implicit construction. We now provide an explicit construction of the wavelet transform. First, observe that Property (4.6) of Theorem 57 applied to $A = \emptyset$ implies that for any $F \in L(\Gamma(\bar{0}))$, $\Psi_\emptyset F = F$. Applying Property (4.7), one obtains for any $F \in L(\bar{\Gamma}_n)$,

$$\Psi_\emptyset F = \Psi_\emptyset M_\emptyset F = M_\emptyset F = \left(\sum_{\pi \in \bar{\Gamma}_n} F(\pi)\right) \delta_{\bar{0}}. \tag{4.14}$$

On the other hand, one has $\phi_A F = F$ for any $A \in \bar{\mathcal{P}}(\llbracket n \rrbracket)$ and $F \in L(\Gamma(A))$, so that by Theorem 57,

$$\Psi_A F \quad = \quad F \; - \sum_{B \in \bar{\mathcal{P}}(A) \setminus \{A\}} \phi_A \Psi_B F. \tag{4.15}$$

We use Eq. (4.14) and (4.15) to construct the wavelet projections $\Psi_B$ by induction. To this purpose, first observe that by Property (4.7) of Theorem 57, one has $\Psi_B F = \Psi_B M_B F$ for any $B \in \bar{\mathcal{P}}(\llbracket n \rrbracket)$ and $F \in L(\bar{\Gamma}_n)$. This justifies the following definition.

**Definition 70** (Alpha coefficients)**.** For $B \in \bar{\mathcal{P}}(\llbracket n \rrbracket)$ and $\pi, \pi' \in \Gamma(B)$, we define the alpha coefficient

$$\alpha_B(\pi, \pi') = \Psi_B \delta_{\pi'}(\pi) \qquad \text{so that} \qquad \Psi_B F(\pi) = \sum_{\pi' \in \Gamma(B)} \alpha_B(\pi, \pi') M_B F(\pi') \qquad \text{for any } F \in L(\bar{\Gamma}_n).$$

The coefficients $\alpha_B(\pi, \pi')$ from Definition 70 entirely characterize the wavelet projections and are easier to handle. Their recursive calculation requires the following lemma. For a ranking $\pi = \pi_1 \dots \pi_k \in \Gamma_n$ and two indexes $1 \le i < j \le k$, we denote by $\pi_{\llbracket i,j \rrbracket}$ the contiguous subword $\pi_i \dots \pi_j$ of $\pi$.

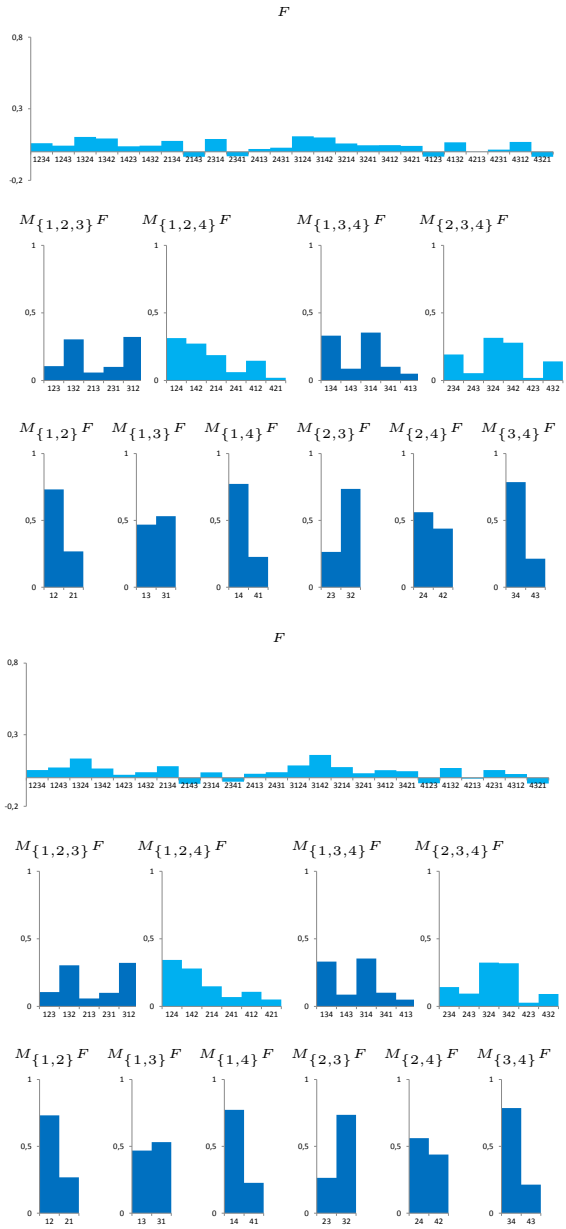



Figure 4.7: Function $F$ and its marginals, with $X_B = 0$ on the left and $X_B$ drawn randomly on the right.

**Lemma 71.** *Let $A \in \bar{\mathcal{P}}([\![n]\!])$ with $|A| = k$ and $\mathbf{X} = (X_B)_{B \in \bar{\mathcal{P}}(A)} \in \mathbb{H}_n$. Then for all $\pi \in \Gamma(A)$,*

$$\sum_{B \in \bar{\mathcal{P}}(A)} \phi_A X_B(\pi) = \frac{1}{k!} X_\emptyset(\bar{0}) + \sum_{1 \le i < j \le k} \frac{1}{(k-j+i)!} X_{c(\pi_{[\![i,j]\!]})}\left(\pi_{[\![i,j]\!]}\right).$$

*Proof.* First, one clearly has $\sum_{B \in \bar{\mathcal{P}}(A)} \phi_A X_B(\pi) = \frac{1}{k!} X_\emptyset(\bar{0}) + \sum_{B \in \mathcal{P}(A)} \phi_A X_B(\pi)$. Now by defi-

nition of operator $\phi_A$, one has for any $B \in \mathcal{P}(A)$

$$\phi_A X_B(\pi) = \sum_{\pi' \in \Gamma(B)} X_B(\pi') \frac{\mathbb{I}\{\pi' \sqsubset \pi\}}{(k - |\pi'| + 1)!} = X_B(\pi_{|B}) \frac{\mathbb{I}\{\pi_{|B} \sqsubset \pi\}}{(k - |B| + 1)!}.$$

Thus only the terms $\phi_A X_B(\pi)$ where $B$ is such that $\pi_{|B}$ is a contiguous subword of $\pi$ are potentially not null in the sum $\sum_{B \in \mathcal{P}(A)} \phi_A X_B(\pi)$. As the contiguous subwords of $\pi$ are all of the form $\pi_{[\![i,j]\!]}$ with $1 \leq i < j \leq k$, this concludes the proof. □

The recursive formula for the coefficients $\alpha_B(\pi, \pi')$ for $\pi, \pi' \in \Gamma(B)$ and $B \in \bar{\mathcal{P}}([\![n]\!])$ is then given by the following theorem.

**Theorem 72** (Recursive formula for the alpha coefficients)**.** *The coefficients* $(\alpha_B(\pi, \pi'))_{\pi, \pi' \in \Gamma(B),\, B \in \mathcal{P}([\![n]\!])}$ *are given by the following recursive formula:*

- $\alpha_\emptyset(\bar{0}, \bar{0}) = 1$
- *for all* $B \in \mathcal{P}([\![n]\!])$ *and* $\pi, \pi' \in \Gamma(B)$,

$$\alpha_B(\pi, \pi') = \mathbb{I}\{\pi = \pi'\} - \frac{1}{|B|!} - \sum_{\substack{1 \leq i < j \leq |B| \\ j - i < |B| - 1}} \frac{1}{(|B| - j + i)!} \alpha_{c(\pi_{[\![i,j]\!]})} \left( \pi_{[\![i,j]\!]}, \pi'_{|c(\pi_{[\![i,j]\!]})} \right).$$

*Proof.* Eq. (4.14) directly implies that $\alpha_\emptyset(\bar{0}, \bar{0}) = 1$. Now, Eq. (4.15) gives for $B \in \mathcal{P}([\![n]\!])$ and $\pi, \pi' \in \Gamma(B)$

$$\Psi_B \delta_{\pi'}(\pi) = \delta_{\pi'}(\pi) \quad - \sum_{B' \in \bar{\mathcal{P}}(B) \setminus \{B\}} \phi_B \Psi_{B'} \delta_{\pi'}(\pi).$$

Combined with Lemma 71, this leads to the desired result. □

*Example* 73. As an example, we provide the matrix $(\alpha_B(\pi, \pi'))_{\pi, \pi' \in \Gamma(B)}$ for $B = \{1, 2\}$ and $B = \{1, 2, 3\}$:

$$\left[\alpha_{\{1,2\}}(\pi, \pi')\right]_{(\pi, \pi')} = \begin{array}{c} \\ 12 \\ 21 \end{array} \begin{array}{c} \begin{array}{cc} 12 & 21 \end{array} \\ \begin{bmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{bmatrix} \end{array}$$

and

$$\left[\alpha_{\{1,2,3\}}(\pi, \pi')\right]_{(\pi, \pi')} = \begin{array}{c} \\ 123 \\ 132 \\ 213 \\ 231 \\ 312 \\ 321 \end{array} \begin{array}{c} \begin{array}{cccccc} 123 & 132 & 213 & 231 & 312 & 321 \end{array} \\ \begin{bmatrix} 1/3 & -1/6 & -1/6 & -1/6 & -1/6 & 1/3 \\ -1/6 & 1/3 & -1/6 & 1/3 & -1/6 & -1/6 \\ -1/6 & -1/6 & 1/3 & -1/6 & 1/3 & -1/6 \\ -1/6 & 1/3 & -1/6 & 1/3 & -1/6 & -1/6 \\ -1/6 & -1/6 & 1/3 & -1/6 & 1/3 & -1/6 \\ 1/3 & -1/6 & -1/6 & -1/6 & -1/6 & 1/3 \end{bmatrix} \end{array}.$$

We conclude this subsection with the following natural property, which will be useful to reduce the complexity of computations.

**Lemma 74.** *Let* $B \in \mathcal{P}([\![n]\!])$ *and* $\tau \in \mathfrak{S}_n$ *a permutation that keeps the order of the items in* $B$*, i.e. such that for all* $b, b' \in B$, $b < b' \Rightarrow \tau(b) < \tau(b')$*. Then for all* $\pi, \pi' \in \Gamma(B)$,

$$\alpha_B(\pi, \pi') = \alpha_{\tau(B)}(\tau(\pi), \tau(\pi')).$$

*Example* 75. Let $B = \{2, 4, 5\}$ and $\tau \in \mathfrak{S}_n$ such that $\tau(2) = 1, \tau(4) = 2$ and $\tau(5) = 3$. Then for $\pi, \pi' \in \Gamma(\{2, 4, 5\})$, $\alpha_{\{2,4,5\}}(\pi, \pi') = \alpha_{\{1,2,3\}}(\tau(\pi), \tau(\pi'))$.

*Proof.* By Definition 70 one has $\Psi_B \delta_{\pi'} = \sum_{\pi \in \Gamma(B)} \alpha_B(\pi, \pi') \delta_\pi$. Applying $T_\tau$ then gives

$$T_\tau \Psi_B \delta_{\pi'} = \sum_{\pi \in \Gamma(B)} \alpha_B(\pi, \pi') \delta_{\tau(\pi)}.$$

On the other hand, Proposition 61 gives

$$T_\tau \Psi_B \delta_{\pi'} = \Psi_{\tau(B)} \delta_{\tau(\pi')} = \sum_{\pi \in \Gamma(\tau(B))} \alpha_{\tau(B)}(\pi, \tau(\pi')) \delta_\pi = \sum_{\pi \in \Gamma(B)} \alpha_{\tau(B)}(\tau(\pi), \tau(\pi')) \delta_{\tau(\pi)}.$$

Identifying the coefficients concludes the proof. □

## 4.3 Fast wavelet transform

The MRA representation would be of little interest without efficient procedures to compute the wavelet transform of a function $F \in L(\bar{\Gamma}_n)$ an the synthesis of an element $\mathbf{X} \in \mathbb{H}_n$. Fortunately, such procedures exist and we now describe them in details. They are directly inspired by the *Fast Wavelet Transform* (FWT) introduced in Mallat (1989). We first recall some background about it.

### 4.3.1 Background on FWT in classic wavelet theory

In classic multiresolution analysis on $l^2(\mathbb{Z})$[1], one is given a scaling basis $(\phi_{j,k})_{j,k \in \mathbb{Z}}$ and a wavelet basis $(\psi_{j,k})_{j,k \in \mathbb{Z}}$, so that any function $f \in l^2(\mathbb{Z})$ decomposes as

$$f = \sum_{k \in \mathbb{Z}} \langle f, \phi_{j_0,k} \rangle \, \phi_{j_0,k} + \sum_{j=j_0}^{+\infty} \sum_{k \in \mathbb{Z}} \langle f, \psi_{j,k} \rangle \, \psi_{j,k}$$

for any $j_0 \in \mathbb{Z}$ (see Mallat, 2008, for the details). The scalars $d_j[k] := \langle f, \psi_{j,k} \rangle$ are the wavelet coefficients and the scalars $a_j[k] := \langle f, \phi_{j,k} \rangle$ are called the approximation coefficients. The fast wavelet transform computes efficiently the wavelet coefficients by exploiting the two following properties of wavelet bases.

- All the wavelet coefficients at scale $j$ can be computed from the approximation coefficients at scale $j$ via a linear operator $h$:

$$d_j[k] = (h a_j)[k] \tag{4.16}$$

- All the approximation coefficients at scale $j$ can be computed from the approximation coefficients at scale $j + 1$ via a linear operator $g$:

$$a_j[k] = (g a_{j+1})[k] \tag{4.17}$$

The operator $g$ computes local averages of the signal and is therefore called a low-pass filter. The operator $h$ computes local differences of the signal and is therefore called a high-pass filter. The FWT then consists in applying recursively these filters in two steps:

[1] $l^2(\mathbb{Z}) = \{f : \mathbb{Z} \to \mathbb{R} \mid \sum_{m \in \mathbb{Z}} f(m)^2 < \infty\}$.

1. Apply the high-pass filter $h$ to $a_j$ to obtain the wavelet coefficients $d_j$
2. Apply the low-pass filter $g$ to $a_j$ to obtain $a_{j-1}$

This procedure is illustrated by Figure 4.8 (the wavelet coefficients are highlighted in blue). It is called "fast" because it computes all the coefficients of a same scale at the same time.

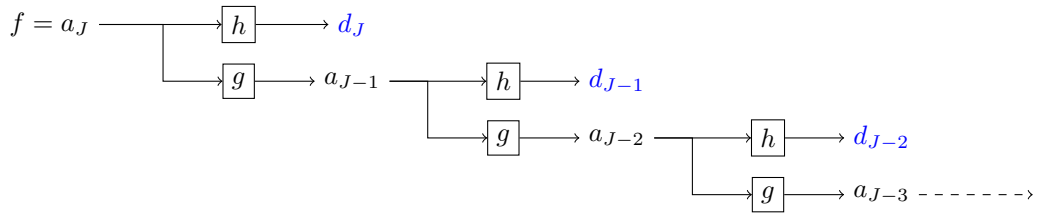
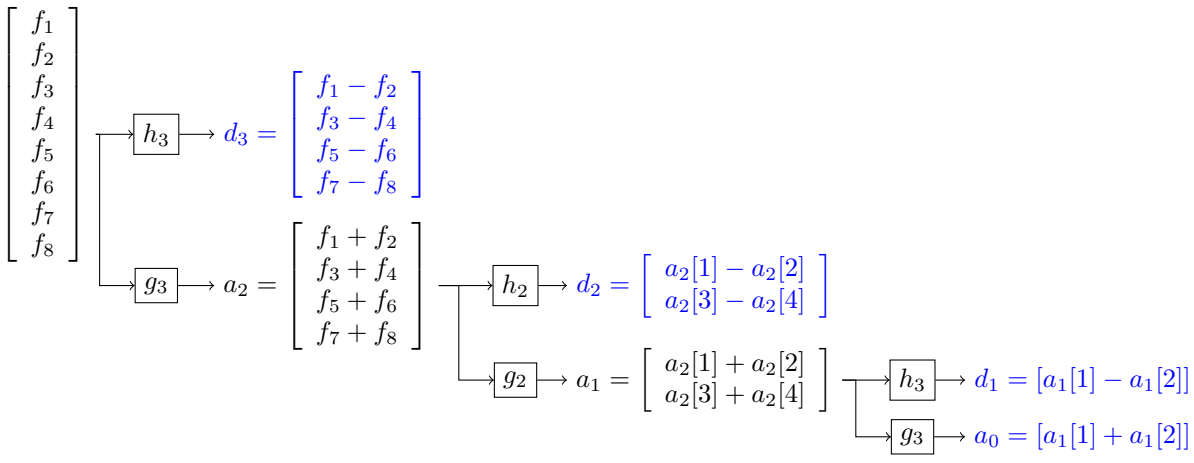



Figure 4.8: Fast Wavelet transform with filter banks

In practice for a function $f \in l^2(\mathbb{Z})$ with finite support, the number of wavelet and approximation coefficients decreases with the scale. The application of the filters $g$ and $h$ at scale $j$ then only involve the operations with the finite vector $a_j$. The implementation of the FWT therefore uses families of filters $(g_j)_j$ and $(h_j)_j$ where $g_j$ and $h_j$ are the operators applied effectively on $a_j$.

*Example* 76 (FWT for the Haar wavelets). The following diagram illustrates the fast Haar wavelet transform of a signal $f = (f_1, \dots, f_8) \in \mathbb{R}^8$.




### 4.3.2 FWT for the MRA representation

We now define the FWT for the MRA representation. We first consider the wavelet transform of a function $F \in L(\Gamma(A))$ with $A \in \mathcal{P}(\llbracket n \rrbracket)$. For $k \in \{2, \dots, |A|\}$ we denote by $\Gamma_A^k := \bigsqcup_{B \subset A,\, |B|=k} \Gamma(B)$ the set of all incomplete rankings of $k$ items of $A$.

- The analogues of the approximation coefficients of $F$ at scale $j \in \{2, \dots, |A|\}$ are the marginals $M_B F$ for $B \subset A$ with $|B| = j$. The vector of approximation coefficients of $F$ at scale $j$ is defined by

$$M^j F = (M_B F(\pi))_{\pi \in \Gamma(B), |B|=j} = (M_{c(\pi)} F(\pi))_{\pi \in \Gamma_A^j} \in \mathbb{R}^{|A|!/(|A|-j)!}. \tag{4.18}$$

- The wavelet coefficients of $F$ at scale $j \in \{2, \dots, |A|\}$ are the wavelet projections $\Psi_B F$ for $B \subset A$ with $|B| = j$. The vector of wavelet coefficients of $F$ at scale $j$ is defined by

$$\Psi^j F = (\Psi_B F(\pi))_{\pi \in \Gamma(B), |B|=j} = (\Psi_{c(\pi)} F(\pi))_{\pi \in \Gamma_A^j} \in \mathbb{R}^{|A|!/(|A|-j)!}. \tag{4.19}$$

Same as in classic wavelet theory, the FWT for the MRA representation also relies on two major relationships between the wavelet and approximation coefficients, analogous to Formulas (4.16) and (4.17). The analogue of Formula (4.17) stems from the properties of the marginal operators. For $\pi = \pi_1 \dots \pi_j \in \Gamma_n$ with $c(\pi) \subsetneq A$, one has

$$M_{c(\pi)} F(\pi) = M_{c(\pi)\cup\{b\}} F(b\pi_1 \dots \pi_j) + M_{c(\pi)\cup\{b\}} F(\pi_1 b \dots \pi_j) + \dots + M_{c(\pi)\cup\{b\}} F(\pi_1 \dots \pi_j b) \tag{4.20}$$

for any $b \in A \setminus c(\pi)$. In addition one has $M_\emptyset F(\bar{0}) = M_{\{a,b\}} F(ab) + M_{\{a,b\}} F(ba)$ for any $a, b \in A$ with $a \neq b$. We therefore define the low-pass filters as follows.

**Definition 77** (Low-pass filters)**.** We define the order 2 low-pass filter $g_A^2 : L(\Gamma_A^2) \to \mathbb{R}\bar{0}$ on $A \in \mathcal{P}(\llbracket n \rrbracket)$ by

$$g_A^2 F(\bar{0}) = F_{\{a,b\}}(ab) + F_{\{a,b\}}(ba) \qquad \text{for any } F \in L(\Gamma_A^2),$$

where $a$ and $b$ are distinct items in $A$ (we take the two smallest by convention). For $j \in \{3, \dots, |A|\}$ we define the order $j$ low-pass filter $g_A^j : L(\Gamma_A^j) \to L(\Gamma_A^{j-1})$ on $A$ by

$$g_A^j F(\pi_1 \dots \pi_{j-1}) = F(b_\pi \pi_1 \dots \pi_{j-1}) + F(\pi_1 b_\pi \dots \pi_{j-1}) + \dots + F(\pi_1 \dots \pi_{j-1} b_\pi)$$

for any $F \in L(\Gamma_A^j)$ and $\pi = \pi_1 \dots \pi_{j-1} \in \Gamma_A^{j-1}$, where $b_\pi$ is any item in $A \setminus c(\pi)$ (we take the smallest by convention).

The high-pass filters are constructed with the alpha coefficients from Definition 70.

**Definition 78** (High-pass filters)**.** Let $A \in \mathcal{P}(\llbracket n \rrbracket)$. For $k \in \{2, \dots, |A|\}$, the high-pass filter on $A$ at scale $j$ is the operator $h_A^j : L(\Gamma_A^j) \to L(\Gamma_A^j)$ defined by

$$h_A^j F(\pi) = \sum_{\pi' \in \Gamma(c(\pi))} \alpha_{c(\pi)}(\pi, \pi') F(\pi') \qquad \text{for any } F \in L(\Gamma_A^j) \text{ and } \pi \in \Gamma_A^j.$$

The analogues of Formulas (4.17) and (4.16) are then given by the following proposition. As it is a direct consequence of Definitions 77 and 78, its proof is left to the reader.

**Proposition 79.** *Let $A \in \mathcal{P}(\llbracket n \rrbracket)$ and $F \in L(\Gamma(A))$.*

- *The wavelet coefficients $\Psi^j F$ of $F$ at scale $j \in \{2, \dots, |A|\}$ can all be computed from the approximation coefficients $M^j F$ at scale $j$ through the high-pass filter $h_A^j$:*

$$\Psi^j F = h_A^j M^j F. \tag{4.21}$$

- *The approximation coefficients $M^j F$ of $F$ at scale $j \in \{2, \dots, |A| - 1\}$ can all be computed from the approximation coefficients $M^{j+1} F$ at scale $j + 1$ through the low-pass filter $g_A^{j+1}$:*

$$M^j F = g_A^{j+1} M^{j+1} F. \tag{4.22}$$

Formulas (4.21) and (4.22) are the respective analogues of Formulas (4.16) and (4.17) in classic wavelet analysis. The FWT for the MRA representation can then be formulated as the FWT in classic wavelet theory: starting from the highest scale, apply recursively the high-pass filter on the approximation coefficients to obtain the wavelet coefficients and the low-pass filter to obtain the approximation coefficients of lower scale. The procedure is formalized in Algorithm 1.

**Algorithm 1** FWT for a function $F \in L(\Gamma(A))$ with $A \in \mathcal{P}(\llbracket n \rrbracket)$

**Require:** $F \in L(\Gamma(A))$ with $A \in \mathcal{P}(\llbracket n \rrbracket)$
  $M^{|A|}F = F$
  **for** $j$ from $|A|$ to 2 **do**
    $\Psi^j F = h_A^j M^j F$
    $M^{j-1}F = g_A^j M^j F$
  **end for**
  **return** $\Psi F = \{M^1 F\} \cup (\Psi^j F)_{2 \leq j \leq |A|}$

*Example* 80 (FWT for the MRA representation). The following diagram illustrates the FWT for a function $F \in L(\Gamma(\llbracket 3 \rrbracket))$. For any $F' \in L(\bar{\Gamma}_n)$ and $\pi \in \bar{\Gamma}_n$, the value $F'(\pi)$ is denoted by $F'_\pi$.

$$\begin{bmatrix} F_{123} \\ F_{132} \\ F_{213} \\ F_{231} \\ F_{312} \\ F_{321} \end{bmatrix} \rightarrow \boxed{h^3_{\llbracket 3 \rrbracket}} \rightarrow \Psi^3_{\llbracket 3 \rrbracket} F = \begin{bmatrix} \sum_{\pi \in \Gamma(\llbracket 3 \rrbracket)} \alpha_{\llbracket 3 \rrbracket}(123, \pi) F_\pi \\ \sum_{\pi \in \Gamma(\llbracket 3 \rrbracket)} \alpha_{\llbracket 3 \rrbracket}(132, \pi) F_\pi \\ \sum_{\pi \in \Gamma(\llbracket 3 \rrbracket)} \alpha_{\llbracket 3 \rrbracket}(213, \pi) F_\pi \\ \sum_{\pi \in \Gamma(\llbracket 3 \rrbracket)} \alpha_{\llbracket 3 \rrbracket}(231, \pi) F_\pi \\ \sum_{\pi \in \Gamma(\llbracket 3 \rrbracket)} \alpha_{\llbracket 3 \rrbracket}(312, \pi) F_\pi \\ \sum_{\pi \in \Gamma(\llbracket 3 \rrbracket)} \alpha_{\llbracket 3 \rrbracket}(321, \pi) F_\pi \end{bmatrix}$$

$$\rightarrow \boxed{g^3_{\llbracket 3 \rrbracket}} \longrightarrow M^2 F = \begin{bmatrix} F_{123} + F_{132} + F_{312} \\ F_{213} + F_{231} + F_{321} \\ F_{132} + F_{123} + F_{213} \\ F_{312} + F_{321} + F_{231} \\ F_{231} + F_{213} + F_{123} \\ F_{321} + F_{312} + F_{231} \end{bmatrix} \rightarrow \boxed{h_2} \rightarrow \Psi^2_{\llbracket 3 \rrbracket} F = \begin{bmatrix} (M^2F)_{12} - (M^2F)_{21} \\ (M^2F)_{21} - (M^2F)_{12} \\ (M^2F)_{13} - (M^2F)_{31} \\ (M^2F)_{31} - (M^2F)_{13} \\ (M^2F)_{23} - (M^2F)_{32} \\ (M^2F)_{32} - (M^2F)_{23} \end{bmatrix}$$

$$\rightarrow \boxed{g_2} \longrightarrow \Psi_\emptyset F = [(M^2F)_{12} + (M^2F)_{21}]$$

Same as the FWT in classic wavelet theory, we call Algorithm 1 a "fast" wavelet transform because it computes all the coefficients of a same scale at the same time. Several differences are worth being pointed out though. We refer the reader to Mallat (2008) for background on classic wavelet theory.

- **Forest structure instead of tree structure.** The classic FWT involves a recursive partitioning of the signal space: at each scale $j$, the vector of approximation coefficients $a_j$ is partitioned into sub-vectors and each sub-vector is averaged to output the approximation coefficients at scale $j-1$. This structure is encoded in the definition of the low-pass filter, Example 76 provides an illustration. The recursive partitioning can be represented by a tree, as shown by Figure 4.9. By contrast, the FWT for the MRA representation follows more a "forest structure", namely the multiscale structure of the marginals represented by Figure 4.4. At scale $j$, each approximation coefficient can be computed as the average of several subsets of approximation coefficients of scale $j+1$, as shown by Equation (4.20). As a consequence, the low-pass filters from Definition 77 are defined up to a convention. They correspond to a certain choice of a spanning tree for the forest structure of the marginals, as illustrated by Figure 4.10.

- **Downsampling.** The FWT in classic wavelet theory more specifically relies on a binary

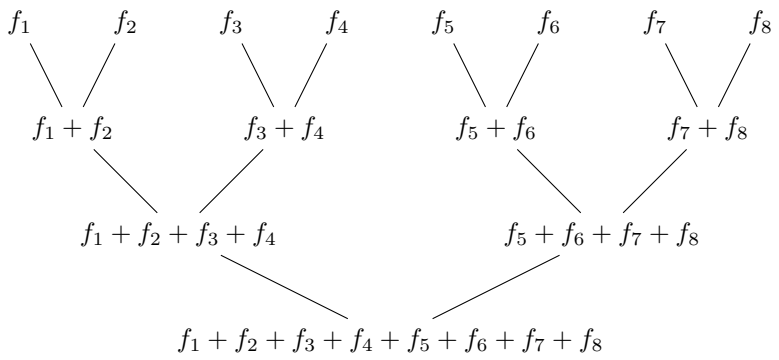



Figure 4.9: Tree structure of the FWT in classic wavelet theory

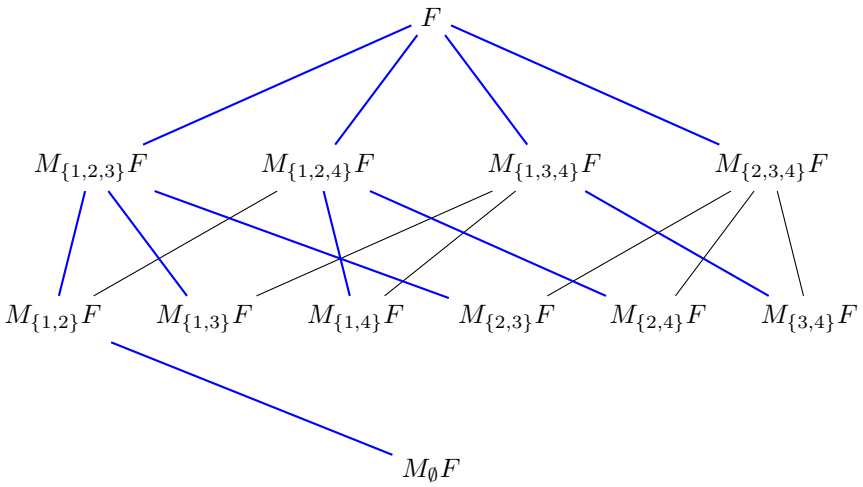



Figure 4.10: Forest structure of the FWT for the MRA representation for $A = [\![4]\!]$. The spanning tree highlighted in blue is the one obtained for $b_\pi = \min A \setminus c(\pi)$ in the Definition 77 of the low-pass filters.

tree structure. At each step, the low-pass filter therefore divides the number of approximation (and thus also wavelet) coefficients by 2. Example 76 provides an illustration. In the MRA representation, the number of approximation and wavelet coefficients of a function $F \in L(\Gamma(A))$ with $A \in \mathcal{P}(\llbracket n \rrbracket)$ at scale $j \in \{2, \dots, |A|\}$ is equal to $|A|!/(|A|-j)!$, as shown by Equations 4.18 and (4.19). Hence at scale $j$, the FWT divides the number of coefficients by $(|A|-j)$.

- **Support of the high-pass filters.** In classic wavelet theory, each wavelet coefficient at scale $j$ is computed from a specific subset of approximation coefficients at scale $j$. Equivalently, each approximation coefficient is involved in the computation of only one wavelet coefficient. As a consequence, the computation of all the wavelet coefficients at scale $j$ can be done in one convolution of the vector $a_j$. The structure of the high-pass filter is a little more complicated in the MRA representation: for a function $F \in L(\Gamma(A))$ with $A \in \mathcal{P}(\llbracket n \rrbracket)$ and a subset $B \in \mathcal{P}(A)$, the computation of each of the wavelet coefficients $\Psi_B F(\pi)$ for $\pi \in \Gamma(B)$ involves all the approximation coefficients $M_B F(\pi')$ for $\pi' \in \Gamma(B)$, by Definition 78 of the high-pass filters. This means that for $j \in \{2, \dots, |A|\}$, the application of the high-pass filter $h_A^j$ requires $j!$ convolutions of the vector $M^j F$.

*Remark* 81 (Further Optimization of the FWT). We point out that the FWT could be further optimized. Indeed for $B \in \mathcal{P}(\llbracket n \rrbracket)$, the space $H_B$ has dimension $d_{|B|}$, whereas the wavelet $\Psi_B F$ projection of a function $F \in L(\bar{\Gamma}_n)$ on $H_B$ is a vector of size $|B|!$. A fully optimized procedure would therefore compute only $d_{|B|}$ scalar coefficients and not $|B|!$. This could be done for instance with the use of a wavelet basis (see Section 4.4 for more details). This direction is left for future work.

The aforementioned differences between the FWT in classic wavelet theory and the FWT for the MRA representation are due to the specific combinatorial structure of the latter. They also stem from the differences between the notions of information localization. In classic multiresolution analysis, the wavelet coefficients are localized in "space" and "scale", where "space" is the very object the signal is defined on. In other words, the metric in this space corresponds to the difference between the indexes of the coordinates: for a signal $f = (f_1, \dots, f_m) \in \mathbb{R}^m$, $f_i$ and $f_{i'}$ corresponds to the values of the function $f$ at points that are separated by a distance of $|i' - i|$. Then at each scale, the coordinates are partitioned recursively into subsets of adjacent coordinates (see Figure 4.9), defining a metric for the scale that is coarser but consistent with the metric of the higher scales. Each wavelet coefficient is thus localized in space and scale because its computation only involves a small number of approximation coefficients that are close with respect the scale.

The notion of information localization in the MRA representation is fundamentally different. The signal is defined on rankings but the wavelet coefficients are localized in "items" and "scale". They thus do not localize components of the signal in the "space of rankings". In other words each wavelet coefficient is not computed from a subset of the signal's coordinates that are "close". Instead, they are computed from subsets of coordinates that lead to the localization properties through the marginal operators that we described at length in the previous subsections.

Algorithm 1 computes the wavelet transform of a function $F \in L(\Gamma(A))$ with $A \in \mathcal{P}(\llbracket n \rrbracket)$. To extend it for any function $F \in L(\bar{\Gamma}_n)$, recall that $\Psi F = \sum_{A \in \mathbf{supp}(F)} \Psi F_A$, where $\mathbf{supp}(F) = \{A \in \bar{\mathcal{P}}(\llbracket n \rrbracket) \mid F_A \neq 0\}$ is the global support of $F$ (see Subsection 4.2.1). We naively extend the FWT by applying Algorithm 1 to each $F_A$ and summing all the wavelet transforms $\Psi F_A$. This procedure is formalized by Algorithm 2.

Algorithm 2 is of course not optimal to compute the wavelet transform of any function $F \in \bar{\Gamma}_n$. Indeed, if there exists $B \in \mathcal{P}(\llbracket n \rrbracket)$ included in at least two subsets of items in $\mathbf{supp}(F)$, then the computation of the wavelet coefficients $\Psi_B F(\pi)$ for $\pi \in \Gamma(B)$ will involve redundant

**Algorithm 2** FWT for a function $F \in L(\bar{\Gamma}_n)$

**Require:** $F \in L(\bar{\Gamma}_n)$
  **for** $A \in \mathbf{supp}(F)$ **do**
    Compute $\Psi F_A$ with Algorithm 1
  **end for**
  **return** $\Psi F = \sum_{A \in \mathbf{supp}(F)} \Psi F_A$

applications of the high-pass filters of scale $|B|$ whereas it requires only one. The definition of an optimal FWT for any function $F \in L(\bar{\Gamma}_n)$ necessitates however to introduce new definitions and notations. For clarity's sake, we leave it to the reader. In addition, we assert that the optimal FWT would still have a complexity of same order of magnitude as the one of Algorithm 2 (see below).

### 4.3.3 Algorithmic complexity

We now turn to the analysis of the complexity of the FWT and related computations. First, the high-pass filters $h_A^j$ are constructed from the alpha coefficients given by Definition 70. They can be computed efficiently once and for all using Theorem 72 and Lemma 74. The following proposition gives an upper bound for the associated complexity. Its proof is left in Appendix.

**Proposition 82** (Complexity of the computation of alpha coefficients)**.** *For $k \in \{2, \dots, n\}$, the computation of all coefficients $\alpha_B(\pi, \pi')$ for $\pi, \pi' \in \Gamma(B)$ and $B \in \bar{\mathcal{P}}(\llbracket n \rrbracket)$ with $|B| \leq k$ has complexity bounded by $(1/2)k^2 k!$.*

Once the alpha coefficients and therefore the high-pass filters are precomputed, one can apply the FWT, the complexity of which is bounded by the following proposition. We recall that the support of a function $F \in L(\bar{\Gamma}_n)$ is defined by $\mathrm{supp}(F) = \{\pi \in \bar{\Gamma}_n \mid F(\pi) \neq 0\}$ whereas its global support is defined by $\mathbf{supp}(F) = \{A \in \mathcal{P}(\llbracket n \rrbracket) \mid F_A \neq 0\}$.

**Proposition 83** (Complexity of the FWT for the MRA representation)**.** *Let $F \in L(\bar{\Gamma}_n)$ and $k = \max\{|A| \mid A \in \boldsymbol{supp}(F)\}$. The complexity of Algorithm 2 applied to $F$ is bounded by*

$$\sum_{A \in \boldsymbol{supp}(F)} [e\,|A|! + |A|(2^{|A|-1} - 1)]|\,\mathrm{supp}(F_A)| \quad \leq \quad [e\,k! + k(2^{k-1} - 1)]|\,\mathrm{supp}(F)|.$$

*Proof.* We first prove the proposition for a function $F \in L(\Gamma(A))$ with $A \in \mathcal{P}(\llbracket n \rrbracket)$. Let $k = |A|$ and $j \in \{2, \dots, k\}$. At scale $j$, Algorithm 1 involves

- the application of the high-pass filter $h_A^j$ on $M^j F$, with complexity equal to

$$\sum_{B \subset A, |B|=j} \sum_{\pi \in \Gamma(B)} |\,\mathrm{supp}(M_B F)| = j! \sum_{B \subset A, |B|=j} |\,\mathrm{supp}(M_B F)|;$$

- the application of the low-pass filter $g_A^j$ on $M^j F$, with complexity bounded by

$$\sum_{\pi \in \Gamma_A^j} \mathbb{I}\{\pi \in \mathrm{supp}(M^j F)\} j = j|\,\mathrm{supp}(M^j F)|.$$

  Indeed, each coefficient $M^j F(\pi)$ for $\pi \in \Gamma_A^j$ is involved in the computation of at most $j$ approximation coefficients of scale $j-1$, namely the approximation coefficients $M^{j-1} F(\pi')$ for $\pi' \subset \pi$ with $|\pi'| = j - 1$.

Now, it is easy to see that for any $B \in \mathcal{P}([\![n]\!])$, $|\operatorname{supp}(M_B F)| \leq |\operatorname{supp}(F)|$. One therefore has $|\operatorname{supp}(M^j F)| = \sum_{B \subset A, |B|=j} |\operatorname{supp}(M_B F)| \leq \binom{k}{j} |\operatorname{supp}(F)|$ and the complexity of Algorithm 1 is bounded by

$$|\operatorname{supp}(F)| \sum_{j=2}^{k} \binom{k}{j} (j! + j)\,.$$

Classic combinatorial calculations then give

$$\sum_{j=2}^{k} \binom{k}{j} j! = \sum_{j=2}^{k} \frac{k!}{(k-j)!} \leq k! \sum_{j=0}^{+\infty} \frac{1}{j!} = e\,k! \qquad \text{and} \qquad \sum_{j=2}^{k} \binom{k}{j} j = k(2^{k-1} - 1).$$

For a function $F \in L(\bar{\Gamma}_n)$, the complexity of Algorithm 2 is then clearly bounded by

$$\sum_{A \in \mathbf{supp}(F)} [e\,|A|! + |A|(2^{|A|-1} - 1)]|\operatorname{supp}(F_A)| \quad \leq \quad [e\,k! + k(2^{k-1} - 1)]|\operatorname{supp}(F)|.$$

□

We finish this subsection with the analysis of the wavelet synthesis. In classic multiresolution analysis, the inverse wavelet transform can be computed with a "dual" procedure of the FWT. In the present context, it happens that the synthesis operator $\phi_A$ involves computations that are not similar to the ones involved in the wavelet transform. The application of Lemma 71 leads however directly to the following bound. The proof is left to the reader

**Proposition 84** (Complexity of the wavelet synthesis)**.** *Let $A \in \mathcal{P}([\![n]\!])$ and $\mathbf{X} \in \mathbb{H}_n$. The computation of $\phi_A \mathbf{X}(\pi)$ can be done with complexity bounded by $\binom{|A|}{2}$ for any $\pi \in \Gamma(A)$, and the computation of $\phi_A \mathbf{X}$ with complexity bounded by $|A|!\binom{|A|}{2}$.*

The complexity bounds of Propositions 82 and 83 can appear a little high at first glance, as they involve powers and factorials. We however point out that the value of the exponent or under the factorial is the size of the subset of items considered. This size is actually small in practical applications typically around 10, and the complexity thus does not explode.

*Remark* 85 (Connection with the Fourier transform). In classic multiresolution analysis, the wavelet transform is connected to the Fourier transform. As we shall see in Chapter 6, it happens that some connections exist too in the present context between the MRA representation and $\mathfrak{S}_n$-based harmonic analysis. The algorithms we introduced in this section do not however use the Fourier transform on $\mathfrak{S}_n$ at all. The design of such procedures would certainly be an interesting direction for future work.

## 4.4 Wavelet basis

As already mentioned, the features of the wavelet transform $\Psi F$ of a function $F \in L(\bar{\Gamma}_n)$ are the vector wavelet projections $\Psi_B F$ for $B \in \bar{\mathcal{P}}([\![n]\!])$. In practice, one may need to decompose a vector $\Psi_B F$ on a basis of the space $H_B$. This section introduces a generative algorithm to explicitly construct such a basis.

Here we adopt an alternative notation for elements of $L(\bar{\Gamma}_n)$: we write them as *chains* (also called *free linear combinations of words*). The function $F = \sum_{\pi \in \bar{\Gamma}_n} F(\pi)\delta_\pi$ is therefore denoted by $F = \sum_{\pi \in \bar{\Gamma}_n} F_\pi \pi$. It does not change anything about the mathematical objects but it makes the calculations easier.

### 4.4.1 Generative algorithm

To define the basis, we use an algorithm adapted from Ragnarsson and Tenner (2011). The latter requires some definitions about cycles and permutations. A cycle on $[\![n]\!]$ is a permutation $\gamma \in \mathfrak{S}_n$ for which there exist $m$ distinct elements $a_1, \dots, a_m \in [\![n]\!]$, with $m \geq 2$, such that $\gamma(a_i) = a_{i+1}$ for $i = 1, \dots, m-1$, $\gamma(a_m) = a_1$, and $\gamma(b) = b$ for all $b \in [\![n]\!] \setminus \{a_1, \dots, a_m\}$. The cycle $\gamma$ is then denoted by $(a_1 \dots a_m)$, its support is the set $\{a_1, \dots, a_m\}$ and its length is $l(\gamma) = m$. For $B \in \mathcal{P}([\![n]\!])$, we denote by $\mathrm{Cycle}(B)$ the set of all cycles with support $B$. It is well known that a permutation $\tau \in \mathfrak{S}_n$ admits a unique decomposition as a product of cycles with distinct supports $\tau = \gamma_1 \dots \gamma_r$ (fixed-points are not represented). This decomposition can though be written in several ways, depending on the order of the cycles and the first element of each cycle.

**Definition 86** (Standard cycle form)**.** A permutation is written in standard cycle form if it is written as a product of disjoint cycles so that the minimum element of a cycle appears at the leftmost letter in that cycle, and the cycles are arranged from left to right in increasing values of minimum letters.

*Example* 87. The permutation $(134)(25)$ is written in standard cycle form, while the alternative representations $(413)(25)$ or $(25)(134)$ are not.

For a permutation $\tau \in \mathfrak{S}_n$, we denote by $\mathrm{cyc}(\tau)$ the number of its cycles, define its support by $\mathrm{supp}(\tau) = \{i \in [\![n]\!] \mid \tau(i) \neq i\}$ and its length by $l(\tau) = |\mathrm{supp}(\tau)|$. These definitions extend the ones of the support and the length for a cycle, and if $\gamma_1 \dots \gamma_{\mathrm{cyc}(\tau)}$ is the cycle decomposition of $\tau$, $l(\tau) = l_1 + \dots + l_{\mathrm{cyc}(\tau)}$. For $B \in \mathcal{P}([\![n]\!])$, we define

$$\mathrm{Der}(B) = \{\tau \in \mathfrak{S}_n \mid \mathrm{supp}(\tau) = B\},$$

and we set by convention $\mathcal{D}_\emptyset = \{id\}$, where $id \in \mathfrak{S}_n$ is the identity permutation on $[\![n]\!]$. If $\tau \in \mathfrak{S}_n$ is a permutation and $B \subset [\![n]\!]$ is a subset such that $\tau(B) = B$ then the restriction of $\tau$ to $B$ is a permutation of $B$, called the induced permutation of $\tau$ on $B$. By definition, a permutation $\tau \in \mathrm{Der}(B)$ induces a fixed-point free permutation, also called a derangement, on $B$. The set $\mathrm{Der}(B)$ is thus the natural embedding of the set of derangements on $B$ in $\mathfrak{S}_n$. In order to state the generative algorithms, one requires some other definitions.

**Definition 88** (Concatenation product)**.** Let $\pi, \pi' \in \bar{\Gamma}_n$ be two injective words. Their concatenation product is then defined by

$$\pi\pi' := \begin{cases} \pi_1 \dots \pi_{|\pi|} \pi'_1 \dots \pi'_{|\pi'|} & \text{if } c(\pi) \cap c(\pi') = \emptyset, \\ 0 & \text{if } c(\pi) \cap c(\pi') \neq \emptyset. \end{cases}$$

It naturally extends to two elements $X, Y \in L(\bar{\Gamma}_n)$ as

$$XY = \sum_{\pi \in \bar{\Gamma}_n} \sum_{\pi' \in \bar{\Gamma}_n} X_\pi Y_{\pi'} \pi\pi'.$$

The algorithm of Ragnarsson and Tenner (2011) computes a basis for the top homology space of the complex of injective words over the field $\mathbb{F}_2 = \mathbb{Z}/2\mathbb{Z}$ of two elements. It uses the operation on $\mathbb{F}_2$-valued chains "$x \diamond y = xy + yx$". In the present setting, we use the following definition.

**Definition 89** (Diamond operator)**.** For $X, Y \in L(\bar{\Gamma}_n)$, we define

$$X \diamond Y = XY - YX.$$

The algorithm of Ragnarsson and Tenner (2011) takes a derangement of $\{1, \dots, k\}$ as input and outputs an element of the top homology space of the complex of injective words. It happens that the same algorithm with the diamond operator of definition 89 maps a derangement of $\{1, \dots, k\}$ to an element of $H_{[\![k]\!]}$, as we shall show in Subsection 4.4.2. More, we extend the algorithm to take a permutation $\tau \in \mathrm{Der}(B)$ as input and output an element $X_\tau$ of the space $H_B$, for any $B \in \mathcal{P}([\![n]\!])$. For clarity's sake, we write the algorithm as a procedure and not in pseudo-code.

**Algorithm 3** Generative algorithm for a basis of $H_B$

Let $B \in \mathcal{P}([\![n]\!])$. The input is a permutation $\tau \in \mathrm{Der}(B)$ written in standard cycle form, and the output is a chain $X_\tau \in H_B$.

| | |
|---|---|
| Step 1. | Between each consecutive pair of letters in each cycle of $\tau$, insert the symbol $\star$. |
| Step 2. | If there are no $\star$ symbols in the string, then HALT. Otherwise, determine which symbol $\star$ has the largest right-hand neighbor. |
| Step 3. | Suppose that the symbol located in Step 2 is between quantities $Q$ and $R$; that is, it appears as $Q \star R$. Then replace $Q \star R$ by $(Q \diamond R)$. |
| Step 4. | GOTO Step 2. |

*Example* 90. Let $A = \{1, 2, 3, 4, 5\}$ and $\tau = (134)(25)$. Algorithm 3 gives the following sequence of steps.

$$\begin{gathered}(1 \star 3 \star 4)(2 \star 5)\\ (1 \star 3 \star 4)(2 \diamond 5)\\ (1 \star (3 \diamond 4))(2 \diamond 5)\\ (1 \diamond (3 \diamond 4))(2 \diamond 5)\end{gathered}$$

Expanding the concatenation and diamond operations, one obtains:

$$\begin{aligned}X_{(134)(25)} &= (1 \diamond (3 \diamond 4))(2 \diamond 5)\\ &= (1 \diamond (34 - 43))(25 - 52)\\ &= (134 - 143 - 341 + 431)(25 - 52)\\ &= 13425 - 13452 - 14325 + 14352 - 34125 + 34152 + 43125 - 43152.\end{aligned}$$

### 4.4.2 Wavelet basis

As announced, the following theorem shows that Algorithm 3 generates a basis for the space $H_B$, for each $B \in \mathcal{P}([\![n]\!])$. For $\tau = id \in \mathfrak{S}_n$, one has $\mathrm{supp}(\tau) = \emptyset$ and we define by convention $X_{id} = \delta_{\bar{0}}$.

**Theorem 91.** *The two following property holds*

1. *For all $\tau \in \mathfrak{S}_n$, $X_\tau \in H_{\mathrm{supp}(\tau)}$.*
2. *For all $B \in \bar{\mathcal{P}}([\![n]\!])$, $(X_\tau)_{Der(B)}$ is a basis of $H_B$.*

*As a consequence, $(X_\tau)_{\tau \in \mathfrak{S}_n}$ is a basis of the feature space $\mathbb{H}_n$ and more generally for any collection $\mathcal{S} \subset \bar{\mathcal{P}}([\![n]\!])$, $(X_\tau)_{\tau \in Der(B),\ B \in \mathcal{S}}$ is a basis of $\mathbb{H}(\mathcal{S})$.*

The proofs of Theorem 91's Properties 1 and 2 are entirely analogous to the ones of lemma 4.3 and theorem 5.2 in Ragnarsson and Tenner (2011). We reproduce the one of Property 1 in Appendix with the notations of the present paper to give some insights. The proof of Property 2 requires by contrast the introduction of several new concepts, hence we let the reader adapt it from Ragnarsson and Tenner (2011). In Table 4.3 we provide a graphical representation of all the elements of the wavelet basis of $H_{[\![k]\!]}$ for $k = 2, 3, 4$.

We now establish some properties about the wavelet chains $X_\tau$. The first result provides some general characterizations.

**Proposition 92.** *Let $\tau \in \mathfrak{S}_n \setminus \{id\}$, $k = |\tau|$ and $r = \mathrm{cyc}(\tau)$. The two following properties hold.*

1. *$X_\tau(\pi) \in \{-1, 0, 1\}$ for all $\pi \in \Gamma(\mathrm{supp}(\tau))$.*
2. *$|\mathrm{supp}(X_\tau)| = 2^{k-r}$.*

*Proof.* The proof is a simple analysis of algorithm 3. For a cycle $\gamma = (a_1 \dots a_k)$, the associated $X_\gamma$ is equal to an expression of the form $a_1 \diamond \cdots \diamond a_k$ with a particular way to put parentheses. When expanded, this expression gives $2^{k-1}$ terms with sign $+$ or $-$ between them. It could happen that some of the terms are the same and thus add or balance. But actually, for $X \in L(\Gamma(A))$ with $A \subset [\![n]\!]$, $1 \leq |A| \leq n-1$ and $b \in [\![n]\!] \setminus A$, $\mathrm{supp}(x \diamond b) = \{\pi b \mid \pi \in \mathrm{supp}(x)\} \sqcup \{b\pi \mid \pi \in \mathrm{supp}(x)\}$. By recursion, we obtain that $|\mathrm{supp}(x_\gamma)| = 2^{k-1}$, meaning also that all the terms in the expanded version of $a_1 \diamond \cdots \diamond a_k$ are different. Furthermore, for $x \in L(\Gamma(A))$ and $y \in L(\Gamma(B))$ with $A, B \subset [\![n]\!]$, $A, B \neq \emptyset$ and $A \cap B = \emptyset$, we have $|\mathrm{supp}(xy)| = |\mathrm{supp}(x)||\mathrm{supp}(y)|$. Now, let $\tau = \gamma_1 \dots \gamma_r$ be a permutation written in standard cycle form, with $\gamma_i = (a_{i,1} \dots a_{i,k_i})$. Then $x_\tau = (a_{1,1} \diamond \cdots \diamond a_{1,k_1}) \dots (a_{r,1} \diamond \cdots \diamond a_{r,k_r})$, and this expression expands in $2^{k_1-1} \dots 2^{k_r-1} = 2^{k-r}$ different terms. This shows both that $|\mathrm{supp}(x_\tau)| = 2^{k-r}$ and that $x_\tau$ takes its values in $\{-1, 0, 1\}$. Applying $\phi_n$ concludes the proof. □

Proposition 92 provides some general intuition about the wavelet basis. In particular, property *1.* is interesting because it means that all the properties of a wavelet function simply depend on the sign of its values and on the combinatorial structure of its support. Notice that both properties of Proposition 92 can be verified on the examples of wavelet basis given in Table 4.3. The second property we establish concerns the relationship between wavelet chains and translation operators.

**Proposition 93.** *Let $\tau \in \mathfrak{S}_n$ and $\tau_0 \in \mathfrak{S}_n$ a permutation that preserves the order of the elements of $\mathrm{supp}(\tau)$, that is if $\mathrm{supp}(\tau) = \{a_1, \dots, a_k\}$ with $a_1 < \cdots < a_k$, then $\tau_0(a_1) < \cdots < \tau_0(a_k)$. Then we have*

$$T_{\tau_0} X_\tau = X_{\tau_0 \tau \tau_0^{-1}}.$$

*Proof.* If $\tau = id$, $\psi_{id}$ is invariant under translations and the equality is trivially verified. We assume $\tau \neq id$, thus $\psi_\tau = \phi_n x_\tau$. By Proposition 61, $T_{\tau_0} \phi_n x_\tau = \phi_n T_{\tau_0} x_\tau$. Let $\gamma_1 \dots \gamma_r$ be the standard cycle form of $\tau$ with $\gamma_i = (a_{i,1} \dots a_{i,k_i})$. Then it is easy to see that $T_{\tau_0} x_\tau$ is the output of Algorithm 3 when taking as input the permutation with cycle form $\gamma'_1 \dots \gamma'_r$ where $\gamma'_i = (\tau_0(a_{i,1}) \dots \tau_0(a_{i,k_i}))$. The order-preserving condition on $\tau_0$ assures that this is a standard cycle form. The proof is concluded by the classic result (or the simple verification) that this is the cycle form of the permutation $\tau_0 \tau \tau_0^{-1}$. □

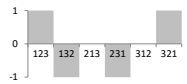
As the same Algorithm 3 is used to generate the basis for each space $H_B$, it is natural that one can obtain the wavelet basis on one space as a translated version from the basis on another space. Algorithm 3 requires however that the input permutation $\tau \in \mathfrak{S}_n$ be written in standard cycle form. This is why Proposition 93 requires a specific assumption. Table 4.4 provides graphical

**Wavelet basis of $H_{\{1,2\}}$**

$X_{(12)}$ $12 - 21$


**Wavelet basis of $H_{\{1,2,3\}}$**

$X_{(123)}$ $123 - 132 - 231 + 321$




$X_{(132)}$ $132 - 213 + 231 - 312$

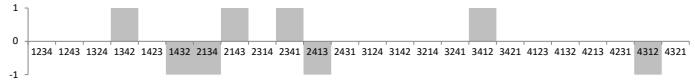

**Wavelet basis of $H_{\{1,2,3,4\}}$**

$X_{(1234)}$ $1234 - 1243 - 1342 + 1432 - 2341 + 2431 + 3421 - 4321$

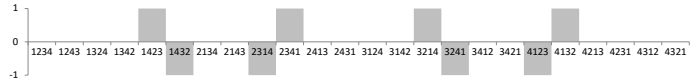

$X_{(1243)}$ $1243 - 1324 + 1342 + 1423 - 2431 + 3241 - 3421 + 4231$


$X_{(1324)}$ $1324 - 1342 - 2413 + 2431 - 3124 + 3142 + 4213 - 4231$

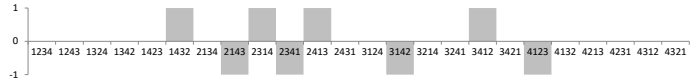

$X_{(1342)}$ $1342 - 1432 - 2134 + 2143 + 2341 - 2431 + 3412 - 4312$



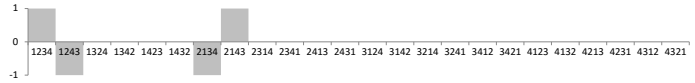

$X_{(1423)}$ $1423 - 1432 - 2314 + 2341 + 3214 - 3241 - 4123 + 4132$



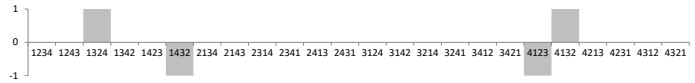

$X_{(1432)}$ $1432 - 2143 + 2314 - 2341 + 2413 - 3142 + 3412 - 4132$




$X_{(12)(34)}$ $1234 - 1243 - 2134 + 2143$




$X_{(13)(24)}$ $1324 - 1342 - 3124 + 3142$




$X_{(14)(23)}$ $1423 - 1432 - 4123 + 4132$

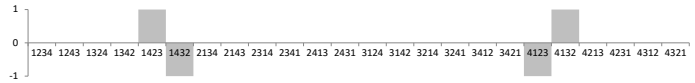



Table 4.3: Wavelet bases of $H_{[\![2]\!]}$, $H_{[\![3]\!]}$ and $H_{[\![4]\!]}$

**Scale 2**

| $X_{(12)}$ | $X_{(13)}$ | $X_{(14)}$ | $X_{(23)}$ | $X_{(24)}$ | $X_{(34)}$ |
|---|---|---|---|---|---|
| $12-21$ | $13-31$ | $14-41$ | $23-32$ | $24-42$ | $34-43$ |
|  |  |  |  |  |  |

**Scale 3**

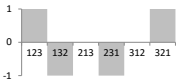
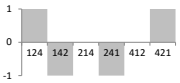
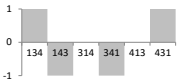
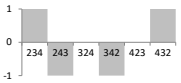
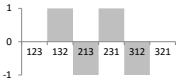
| $X_{(123)}$ | $X_{(124)}$ | $X_{(134)}$ | $X_{(234)}$ |
|---|---|---|---|
| $123-132-231+321$ | $124-142-241+421$ | $134-143-341+431$ | $234-243-342+432$ |
|  |  |  |  |
| $X_{(132)}$ | $X_{(142)}$ | $X_{(143)}$ | $X_{(243)}$ |
| $132-213+231-312$ | $142-214+241-412$ | $143-314+341-413$ | $243-324+341-423$ |
|  |  |  |  |

Table 4.4: Wavelet bases of each $H_B$ for $B \in \mathcal{P}(\llbracket 4 \rrbracket) \setminus \llbracket 4 \rrbracket$

representations of the wavelet bases of the spaces $H_B$ for $B \subset \llbracket 4 \rrbracket$ at scales $|B| = 2, 3$. One can easily verify Proposition 93 on these examples.

One can also be interested in using a wavelet basis for a signal space $L(\Gamma(A))$ for $A \in \mathcal{P}(\llbracket n \rrbracket)$, that would refine the multiresolution decomposition $L(\Gamma(A)) = \bigoplus_{B \in \bar{\mathcal{P}}(A)} \phi_A(H_B)$ given by Theorem 47. As the latter shows that $\phi_A$ is injective on $H_B$ for all $B \in \bar{\mathcal{P}}(A)$, the following result is a direct consequence of Theorem 91. Its proof is left to the reader.

**Corollary 94.** *For all $A \in \mathcal{P}(\llbracket n \rrbracket)$, $(\phi_A X_\tau)_{\tau \in Der(B),\ B \in \bar{\mathcal{P}}(A)}$ is a basis of $L(\Gamma(A))$ consistent with the multiresolution decomposition. In particular, $(\phi_{\llbracket n \rrbracket} X_\tau)_{\tau \in \mathfrak{S}_n}$ is a basis of $L(\mathfrak{S}_n)$ consistent with the multiresolution decomposition.*

Table 4.5 provides a graphical representation for the elements of the wavelet basis of $L(\mathfrak{S}_4)$ at scales 0, 2 and 3. Graphical representations form the elements of scale 4 are given in Table 4.3.

*Remark* 95 (Non orthogonality of the wavelet basis). We point out that neither the basis $(X_\tau)_{\tau \in \mathrm{Der}(B)}$ of $H_B$ for $B \in \mathcal{P}(\llbracket n \rrbracket)$ nor the basis $(\phi_A X_\tau)_{\tau \in \mathrm{Der}(B),\ B \in \bar{\mathcal{P}}(A)}$ of $L(\Gamma(A))$ for $A \in \mathcal{P}(\llbracket n \rrbracket)$ is orthogonal. One has for instance $\left\langle X_{(123)}, X_{(132)} \right\rangle = -2$.

### 4.4.3 Wavelet coefficients

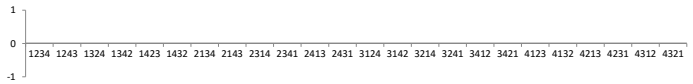
We now refine the wavelet projections by decomposing them in the wavelet basis.

**Definition 96** (Wavelet coefficients)**.** For any $F \in L(\bar{\Gamma}_n)$ and $B \in \bar{\mathcal{P}}(\llbracket n \rrbracket)$, we define the *wavelet coefficients* $(c_\tau(F))_{\tau \in \mathrm{Der}(B)}$ as the coefficients of the wavelet projection $\Psi_B F$ in the wavelet basis $(X_\tau)_{\tau \in \mathrm{Der}(B)}$ of $H_B$. In other words, $(c_\tau(F))_{\tau \in \mathrm{Der}(B)}$ are the only scalars such that

$$\Psi_B F = \sum_{\tau \in \mathrm{Der}(B)} c_\tau(F) X_\tau.$$

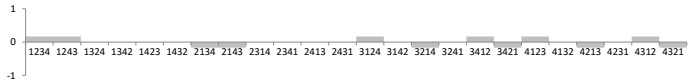
**Scale 0**

$\phi_{[\![4]\!]} X_{id}$



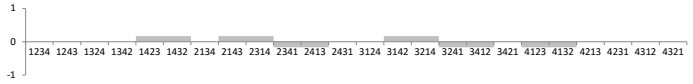

**Scale 2**

$\phi_{[\![4]\!]} X_{(12)}$

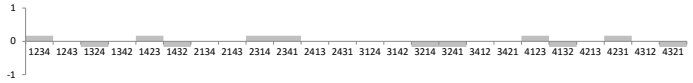



$\phi_{[\![4]\!]} X_{(13)}$

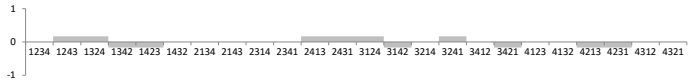

$\phi_{[\![4]\!]} X_{(14)}$



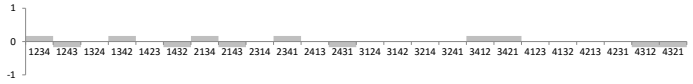

$\phi_{[\![4]\!]} X_{(23)}$




$\phi_{[\![4]\!]} X_{(24)}$

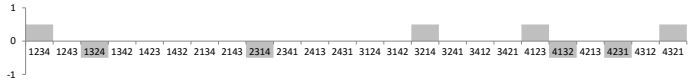



$\phi_{[\![4]\!]} X_{(34)}$

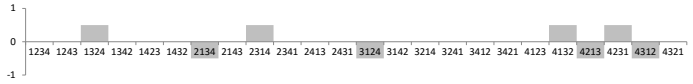



**Scale 3**

$\phi_{[\![4]\!]} X_{(123)}$

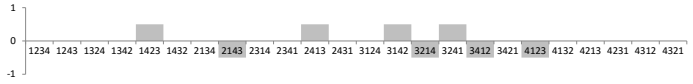



$\phi_{[\![4]\!]} X_{(132)}$

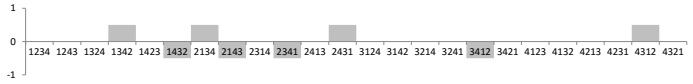



$\phi_{[\![4]\!]} X_{(124)}$

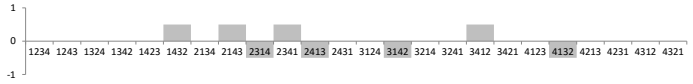

$\phi_{[\![4]\!]} X_{(142)}$



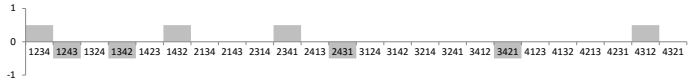

$\phi_{[\![4]\!]} X_{(134)}$




$\phi_{[\![4]\!]} X_{(143)}$




$\phi_{[\![4]\!]} X_{(234)}$




$\phi_{[\![4]\!]} X_{(243)}$


Table 4.5: Elements of the wavelet basis of $L(\mathfrak{S}_4)$ at scales 0, 2 and 3

Theorem 57 naturally reformulates with wavelet coefficients.

**Corollary 97** (Fundamental properties of the wavelet coefficients)**.** *Let $A \in \bar{\mathcal{P}}([\![n]\!])$ and $F \in L(\Gamma(A))$. The wavelet coefficients satisfy the following properties.*

- $(c_\tau(F))_{\tau \in Der(B), B \in \bar{\mathcal{P}}(A)}$ *is the unique element in* $\mathbb{R}^{|A|!}$ *such that*

$$F = \sum_{B \in \bar{\mathcal{P}}(A)} \sum_{\tau \in Der(B)} c_\tau(F) \phi_A X_\tau. \tag{4.23}$$

- *For any $A' \in \bar{\mathcal{P}}(A)$, $B \in \bar{\mathcal{P}}(A')$ and $\tau \in Der(B)$,*

$$c_\tau(M_{A'} F) = c_\tau(F). \tag{4.24}$$

As the wavelet basis is not orthogonal, the wavelet coefficients must be calculated by inverting a linear system. More specifically, for any $B \in \mathcal{P}([\![n]\!])$ and $F \in L(\bar{\Gamma}_n)$, the wavelet coefficients $(c_\tau(F))_{\tau \in \mathrm{Der}(B)}$ are the solutions of the system of equations

$$\sum_{\tau \in \mathrm{Der}(B)} X_\tau(\pi) c_\tau(F) = \Psi_B F(\pi) \tag{4.25}$$

for $\pi \in \Gamma(B)$. For instance for $B = \{1, 2\} \subset [\![n]\!]$, System (4.25) writes as

$$\begin{cases} X_{(12)}(12) c_{(12)}(F) = \Psi_{\{1,2\}} F(12) \\ X_{(12)}(21) c_{(12)}(F) = \Psi_{\{1,2\}} F(21) \end{cases}$$

which gives, by Algorithm 3 and Example 73,

$$c_{(12)}(F) = \Psi_{\{1,2\}} F(12) = \frac{1}{2} \left( M_{\{1,2\}} F(12) - M_{\{1,2\}} F(21) \right).$$

For $B = \{1, 2, 3\}$, System (4.25) writes as

$$\begin{cases} X_{(123)}(123) c_{(123)}(F) + X_{(132)}(123) c_{(132)}(F) = \Psi_{\{1,2,3\}} F(123) \\ X_{(123)}(132) c_{(123)}(F) + X_{(132)}(132) c_{(132)}(F) = \Psi_{\{1,2,3\}} F(132) \\ X_{(123)}(213) c_{(123)}(F) + X_{(132)}(213) c_{(132)}(F) = \Psi_{\{1,2,3\}} F(213) \\ X_{(123)}(231) c_{(123)}(F) + X_{(132)}(231) c_{(132)}(F) = \Psi_{\{1,2,3\}} F(231) \\ X_{(123)}(312) c_{(123)}(F) + X_{(132)}(312) c_{(132)}(F) = \Psi_{\{1,2,3\}} F(312) \\ X_{(123)}(321) c_{(123)}(F) + X_{(132)}(321) c_{(132)}(F) = \Psi_{\{1,2,3\}} F(321) \end{cases}$$

which gives, by Algorithm 3 and Example 73,

$$\begin{cases} c_{(123)}(F) = \Psi_{\{1,2,3\}} F(123) = \frac{1}{6} \left( 2F_{123} - F_{132} - F_{213} - F_{231} - F_{312} + 2F_{321} \right) \\ c_{(132)}(F) = \Psi_{\{1,2,3\}} F(213) = \frac{1}{6} \left( -F_{123} - F_{132} + 2F_{213} - F_{231} + 2F_{312} - F_{321} \right), \end{cases}$$

where $F_\pi$ is a short notation for $M_{\{1,2,3\}} F(\pi)$ for $\pi \in \Gamma(\{1, 2, 3\})$. Solving System (4.25) in general is however much harder and finding an explicit formula for its solutions or designing an efficient algorithm that can generate them would certainly be an interesting direction for future research.

Fortunately, the wavelet basis is already useful in practice even without the coefficients. It enables indeed to obtain an element of $H_B$ for any $B \in \mathcal{P}([\![n]\!])$ as the linear combination $\sum_{\tau \in \mathrm{Der}(B)} c_\tau X_\tau$ for any collection $(c_\tau)_{\tau \in \mathrm{Der}(B)}$. This could be used for instance to design statistical procedures.

*Remark* 98 (Connection with Lyndon words, Lie algebras and Hopf monoids). In this remark we explicit an interesting connection with some other algebraic objects. Let $\gamma = (a_1 \dots a_n) \in \text{Cycle}(\llbracket n \rrbracket)$ be a cycle of support $\llbracket n \rrbracket$ written in standard cycle form, so that $a_1 = 1$. The expression $a_1 \dots a_n$ is thus an injective word on $\llbracket n \rrbracket$ that starts with 1. It is therefore a *Lyndon word* (Chen et al., 1958). Now, it happens that Algorithm 3 applied on $\gamma$ to output $X_\gamma$ is exactly equivalent to the *standard bracketing* (see for instance Diaconis et al., 2014) of $a_1 \dots a_n$ which therefore outputs the same chain. Applying a standard result from Lothaire (1983); Reutenauer (1993), this means in particular that $\{X_\gamma\}_{\gamma \in \text{Cycle}(\llbracket n \rrbracket)}$ spans the $n^{th}$ homogenous component of the *free Lie algebra* over $\llbracket n \rrbracket$ and is its canonical basis. Another connection exists with the construction of Aguiar and Lauve (2011): the basis $\{X_\tau\}_{\tau \in \text{Der}(\llbracket n \rrbracket)}$ is exactly the same as the basis constructed for what the authors call "the Hopf kernel of the canonical morphism of Hopf monoids between the species of linear orders and the exponential species" (see part 5.3). These connections may bring new insights or lead to new results for the MRA framework.

# Chapter 5

# Application to the statistical analysis of incomplete rankings

This chapter is about the application of the MRA representation to the statistical analysis of incomplete rankings. First we define in Section 5.1 a general framework fitted for many statistical tasks. Then we develop its application to the estimation of the marginals of a ranking model in Section 5.2 and to the prediction of rankings on subsets of elements in 5.3.

## Contents



## 5.1 General MRA framework for the statistical analysis of incomplete rankings

We first describe a general framework to apply the MRA representation to the statistical analysis of incomplete rankings, in the setting defined in Section 3.1. Here and throughout the chapter we consider a dataset $\mathcal{D}_N = ((\mathbf{A}_1, \Pi^{(1)}), \dots, (\mathbf{A}_N, \Pi^{(N)}))$ drawn IID from the process (3.3) with $p$ an unknown ranking model and $\nu$ a known probability distribution over $\mathcal{P}(\llbracket n \rrbracket)$ called the observation design.

For $A \in \mathcal{P}(\llbracket n \rrbracket)$, let $\widehat{I}_A = \{1 \leq i \leq N \mid \mathbf{A}_i = A\}$ be the set of indexes $i$ such that $\mathbf{A}_i = A$ and let $\widehat{N}_A = |\widehat{I}_A|$ be the number of times the subset $A$ was observed. The *empirical observation design* is the empirical probability distribution $\widehat{\nu}$ over $\mathcal{P}(\llbracket n \rrbracket)$ defined by $\widehat{\nu}(A) = \widehat{N}_A/N$. Its support is the set $\widehat{\mathcal{A}}_N = \mathrm{supp}(\widehat{\nu}) = \{A \in \mathcal{P}(\llbracket n \rrbracket) \mid \widehat{N}_A > 0\}$ of all observed subsets in $\mathcal{D}_N$. It is necessarily included in the support $\mathcal{A}$ of the observation design $\nu$. We denoted by $\mathcal{B}_N^\nu$ the $\sigma$-algebra generated by $\widehat{\nu}$. By construction $\widehat{\mathcal{A}}_N$, $\widehat{I}_A$ and $\widehat{N}_A$ are $\mathcal{B}_N^\nu$-measurable for all $A \in \mathcal{P}(\llbracket n \rrbracket)$. Unless otherwise specified, the expectation sign $\mathbb{E}$ in this chapter denotes the expectation taken with respect to the drawing of $\mathcal{D}_N$.

### 5.1.1 Identifiability issues

In several applications mentioned in Section 2.3, the goal is to recover a certain target part of $p$. In the context of full ranking analysis, one observes drawings of a random permutation $\Sigma$ that provide a direct access to global information about $p$. The task is then to best approximate the target part of $p$ from global information about $p$. In the context of incomplete ranking analysis, the target part of $p$ must be recovered from the observation of drawings of a random couple $(\mathbf{A}, \Pi)$ drawn from the process (3.3). This brings an additional difficulty as information about $p$ is then censored by the probability distribution $\nu$. One must therefore deal with two types of uncertainty:

1. Remove the noise from the observation process (3.3) to gain access to information about $p$.
2. Recover the target part of $p$ from the accessible part of information about $p$.

By the law of large numbers, it is obvious that the (asymptotically) accessible part of information about $p$ (as $N$ grows to infinity) are the marginals $P_A$ for observable subsets of items $A$, that is to say subsets of items in the observation design $\mathcal{A}$. The second problem then boils down to recover the target part of $p$ from the knowledge of the marginals $(P_A)_{A \in \mathcal{A}}$.

Depending on the target part and the observation design $\mathcal{A}$, this task can require a structural assumption on $p$. Suppose for instance that one seeks to recover the full ranking model $p$ from the observation of pairwise comparisons only. In other words, with an observation design $\mathcal{A}$ included in the set of pairs of $\llbracket n \rrbracket$. Each pairwise marginal $P_{\{a,b\}}$ for $\{a, b\} \subset \llbracket n \rrbracket$ being a probability distribution on a set with two elements, it is characterized by one parameter. The number of accessible parameters is therefore at most $\binom{n}{2}$, whereas characterizing the full ranking model $p$ requires $n! - 1$ parameters. This task thus requires to stipulate an additional structural assumption on $p$, so that $p$ becomes identifiable from the knowledge of its pairwise marginals only.

In a general context, we consider the following question: without any structural assumption, what part of $p$ can be recovered from the knowledge of the marginals $(P_A)_{A \in \mathcal{A}}$? The following theorem provides the answer. Being a direct consequence of Theorem 68, its proof is left to the reader.

**Theorem 99** (Identifiable parameters)**.** *The knowledge of $(P_A)_{A \in \mathcal{A}}$ characterizes the component*

$$(\Psi_B p)_{B \in \bar{\mathcal{P}}(\mathcal{A})} \in \mathbb{H}(\bar{\mathcal{P}}(\mathcal{A}))$$

*of the ranking model $p$. In particular, it has a number of degrees of freedom equal to* $\dim \mathbb{H}(\bar{\mathcal{P}}(\mathcal{A})) = \sum_{B \in \bar{\mathcal{P}}(\mathcal{A})} d_{|B|}$.

Through Theorem 99, the MRA representation allows to quantify the part of $p$ that is identifiable without any structural assumption in the statistical setting introduced in Section 3.1. This justifies the general method we introduce for the statistical analysis of incomplete rankings.

### 5.1.2 General method for the statistical analysis of incomplete rankings

The MRA framework we now introduce is performed in two steps, one to perform each of the two tasks mentioned in the previous Subsection.

**Definition 100** (MRA framework)**.** The MRA framework for the statistical analysis of incomplete rankings is described by the following general procedure.

1. Construct from the dataset $\mathcal{D}_N$ the *wavelet empirical estimator* $\widehat{\mathbf{X}} \in \mathbb{H}(\bar{\mathcal{P}}(\mathcal{A}))$ defined for each $B \in \bar{\mathcal{P}}(A)$ as the simple average of the wavelet projections of the $\delta_{\Pi^{(i)}}$:

$$\widehat{X}_B = \frac{1}{\sum_{A \in \mathcal{Q}(B)} \widehat{N}_A} \sum_{i=1}^{N} \Psi_B \delta_{\Pi^{(i)}} \tag{5.1}$$

   where we recall that $\mathcal{Q}(B) = \{A \in \mathcal{P}(\llbracket n \rrbracket) \mid B \subset A\}$ and that $\Psi_B \delta_\pi = 0$ if $B \not\subset c(\pi)$ by construction. By convention, $\widehat{X}_B = 0$ if $\sum_{A \in \mathcal{Q}(B)} \widehat{N}_A = 0$. As shown in Subsection 5.1.3, $\widehat{\mathbf{X}}$ is an unbiased estimator of the accessible component $(\Psi_B p)_{B \in \bar{\mathcal{P}}(\mathcal{A})}$ of $p$.

2. Perform the task related to the considered application in the feature space $\mathbb{H}_n$ using $\widehat{\mathbf{X}}$ as empirical distribution.

Beyond this decomposition in two steps, the major novelty of the MRA framework is to offer the possibility to perform the analysis of the data in the feature space $\mathbb{H}_n$. This is a radical change from existing approaches that all rely on the construction of a ranking model $\widehat{p}_N$ over $\mathfrak{S}_n$ (see Subsection 3.1.5). Subsections 5.1.3 and 5.1.4 respectively show how this method allows to overcome the statistical and computational challenges.

### 5.1.3 Overcoming the statistical challenge

We now describe the advantages of the MRA framework for the statistical analysis of incomplete rankings. First, $\widehat{\mathbf{X}}$ is an unbiased estimator of $(\Psi_B p)_{B \in \bar{\mathcal{P}}(A)}$.

**Proposition 101** (Expectation of the wavelet empirical estimator)**.** *For all* $B \in \bar{\mathcal{P}}(\mathcal{A})$*,*

$$\mathbb{E}\left[\widehat{X}_B\right] = \Psi_B p.$$

*Proof.* Let $B \in \bar{\mathcal{P}}(\mathcal{A})$. Recalling that $\mathcal{B}_N^\nu$ is the $\sigma$-algebra generated by $\widehat{\nu}$, one has by definition

$$\begin{aligned}
\mathbb{E}\left[\widehat{X}_B\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{1}{\sum_{A \in \mathcal{Q}(B)} \widehat{N}_A} \sum_{i=1}^{N} \mathbb{I}\{B \subset \mathbf{A}_i\} \Psi_B \delta_{\Pi^{(i)}} \middle| \mathcal{B}_N^\nu\right]\right] \\
&= \mathbb{E}\left[\frac{1}{\sum_{A \in \mathcal{Q}(B)} \widehat{N}_A} \sum_{i=1}^{N} \mathbb{I}\{B \subset \mathbf{A}_i\} \mathbb{E}\left[\Psi_B \delta_{\Pi^{(i)}} \middle| \mathcal{B}_N^\nu\right]\right].
\end{aligned}$$

Now, reformulation (3.4) of the statistical process (3.3) ensures that for each $i \in \{1, \ldots, N\}$, $\Pi^{(i)}$ has the same law as $\Sigma^{(i)}_{|\mathbf{A}_i}$, where $\Sigma^{(1)}, \ldots, \Sigma^{(N)}$ are random permutations drawn IID from $p$. We recall in addition that for any permutation $\sigma \in \mathfrak{S}_n$ and any subset $A \in \mathcal{P}(\llbracket n \rrbracket)$ with $B \subset A$, one has $\Psi_B \delta_{\sigma_{|A}} = \Psi_B \delta_\sigma$ by Property (4.7) of Theorem 57. One therefore has

$$\mathbb{E}\left[\Psi_B \delta_{\Pi^{(i)}} \middle| \mathcal{B}_N^\nu\right] = \mathbb{E}\left[\Psi_B \delta_{\Sigma^{(i)}_{|\mathbf{A}_i}} \middle| \mathcal{B}_N^\nu\right] = \mathbb{E}\left[\Psi_B \delta_{\Sigma^{(i)}} \middle| \mathcal{B}_N^\nu\right] = \mathbb{E}\left[\Psi_B \delta_{\Sigma^{(i)}}\right] = \Psi_B \mathbb{E}\left[\delta_\Sigma\right] = \Psi_B p.$$

This concludes the proof. □

Proposition 101 ensures that $\widehat{\mathbf{X}}$ is a good representative of the accessible part $(\Psi_B p)_{B\in\bar{\mathcal{P}}(\mathcal{A})}$ of the ranking model $p$, whatever it is. This advantage is to be compared to existing methods:

- Methods based on parametric models are necessarily biased when the ranking model does not satisfy the structural assumption.
- Methods that identify an incomplete ranking with the set of its linear extensions are fundamentally biased by the censoring process $\nu$, as shown in Subsection 3.1.5.

In a sense, one can say that the MRA framework allows to remove the noise due to the censoring process $\nu$ whatever the ranking model $p$.

The other statistical advantage of the MRA framework is that it allows to fully exploit the consistency assumption $(*)$. As explained in Subsection 3.1.4, the consistency assumption induces two rules to transfer information between subsets of items $A, B \in \mathcal{P}(\llbracket n \rrbracket)$ with $B \subset A$: information is transferred from $A$ to $B$ through the marginal operator $M_B$, and information is transferred from $B$ to $A$ as the constraint that $P_A$ must satisfy $M_B P_A = P_B$. By Theorem 68, this constraint is equivalent to $\Psi_{B'} P_A = \Psi_{B'} P_B$ for all $B' \in \bar{\mathcal{P}}(B)$. The second rule can thus be reformulated as: information is transferred from $B$ to $A$ through the operators $(\Psi_{B'})_{B'\in\bar{\mathcal{P}}(B)}$. In other words, the MRA representation enables to quantify the amount of information in the constraints imposed by the consistency assumption. The wavelet empirical estimator $\widehat{\mathbf{X}}$ therefore naturally exploits more information than other empirical estimators, as illustrated by the following comparison.

- **Naive empirical estimator.** For an observed subset $A$ ($\widehat{N}_A > 0$), we recall that the naive empirical estimator is defined in (3.7) by

$$\widehat{P_A} = \frac{1}{\widehat{N}_A} \sum_{i=1}^{N} \mathbb{I}\{A = \mathbf{A}_i\} \delta_{\Pi^{(i)}}.$$

  The $\widehat{P_A}$'s are two-by-two independent. Each $\widehat{P_A}$ consolidates information on $A$ but no information is transferred between subsets. In other words, the naive empirical estimator does not exploit the consistency assumption at all. For instance if rankings are observed on $\{1, 2\}$ and $\{1, 2, 3\}$, neither information is transferred from $\{1, 2, 3\}$ to $\{1, 2\}$ nor in the other way round.

- **Marginal-based empirical estimator.** For a subset $B \in \bar{\mathcal{P}}(\llbracket n \rrbracket)$ included in at least one observed subset ($\sum_{A\in\mathcal{Q}(B)} \widehat{N}_A > 0$), we define the marginal-based empirical estimator by

$$\tilde{P_B} = \frac{1}{\sum_{A\in\mathcal{Q}(B)} \widehat{N}_A} \sum_{i=1}^{N} \mathbb{I}\{B \subset \mathbf{A}_i\} M_B \delta_{\Pi^{(i)}}.$$

  The marginal-based empirical estimator exploits the consistency assumption but only in one sense, from a subset of item $A$ to its subsets $B \in \bar{\mathcal{P}}(\llbracket n \rrbracket)$. For instance if rankings are observed on $\{1, 2\}$ and $\{1, 2, 3\}$, information is transferred from $\{1, 2, 3\}$ to $\{1, 2\}$ but not in the other way round.

- **Wavelet empirical estimator.** For a subset $B \in \bar{\mathcal{P}}(\llbracket n \rrbracket)$ included in at least one observed subset ($\sum_{A\in\mathcal{Q}(B)} \widehat{N}_A > 0$), we recall that the wavelet empirical estimator is defined by

$$\widehat{X}_B = \frac{1}{\sum_{A\in\mathcal{Q}(B)} \widehat{N}_A} \sum_{i=1}^{N} \mathbb{I}\{B \subset \mathbf{A}_i\} \Psi_B \delta_{\Pi^{(i)}}.$$

Thanks to the wavelet transform, the wavelet empirical estimator fully exploits the consistency assumption. For instance if rankings are observed on $\{1,2\}$ and $\{1,2,3\}$, information is transferred from $\{1,2,3\}$ to $\{1,2\}$ and in the other way round.

### 5.1.4 Overcoming the computational challenge

The following proposition gives a theoretical bound on the complexity of the computation of the wavelet empirical estimator $\widehat{\mathbf{X}}$.

**Proposition 102** (Complexity of the computation of the wavelet empirical estimator)**.** *Let* $K = \max_{A\in\mathcal{A}} |A|$*. The complexity of the computation of* $\widehat{\mathbf{X}}$ *is bounded by*

$$[e\,K! + (K+4)2^{K-1}] \min\left(N, \sum_{A\in\mathcal{A}} |A|!\right).$$

*Proof.* Defining the function $\widehat{F}_N = \sum_{i=1}^N \delta_{\Pi^{(i)}}$ and the scalars $\widehat{Z}_{N,B} = \sum_{A\in\mathcal{Q}(B)} \widehat{N}_A$, one has for any $B \in \bar{\mathcal{P}}(\mathcal{A})$,

$$\widehat{X}_B = \frac{1}{\widehat{Z}_{N,B}} \Psi_B \widehat{F}_N.$$

The computation of $\widehat{\mathbf{X}}$ can thus be decomposed into three steps:

1. Computation of $\widehat{F}_N$ and $(\widehat{Z}_{N,B})_{B\in\bar{\mathcal{P}}(\mathcal{A})}$: this is performed in one loop over the dataset with complexity bounded by
$$\sum_{\pi\in\mathrm{supp}(\widehat{F}_N)} |\bar{\mathcal{P}}(c(\pi))| \leq 2^K |\mathrm{supp}(\widehat{F}_N)|.$$
2. Computation of $\Psi\widehat{F}_N$: this is performed using Algorithm 2. By Proposition 83, its complexity is bounded by
$$[e\,K! + K2^{K-1}]|\mathrm{supp}(\widehat{F}_N)|.$$
3. Division of $\Psi_B\widehat{F}_N$ by $\widehat{Z}_{N,B}$ for each $B \in \bar{\mathcal{P}}(\mathcal{A})$ such that $\widehat{Z}_{N,B} \neq 0$: this is performed in one loop over the subsets $B$ with $\widehat{Z}_{N,B} > 0$ with complexity bounded by
$$|\bar{\mathcal{P}}(\mathbf{supp}(\widehat{F}_N))| \leq 2^K |\mathrm{supp}(\widehat{F}_N)|.$$

To conclude the proof, notice that $|\mathrm{supp}(\widehat{F}_N)|$ is exactly the number of parameters required to store the dataset $\mathcal{D}_N$. Lemma 17 therefore ensures that it is bounded by $\min(N, \sum_{A\in\mathcal{A}} |A|!)$. □

Although the bound in Proposition 102 is not small, it is sufficient to ensure that the computation of the wavelet empirical estimator is tractable in common situations. In practical applications indeed, the number of items $n$ can be large, say around $10^4$, but the parameter $K$, which represents the maximal size of an observed ranking, is fairly small, typically less than 10. The factor $[e\,K! + K2^{K-1}]$ then does not represent too much of an issue. On the other hand, the term $\min\left(N, \sum_{A\in\mathcal{A}} |A|!\right)$ is smaller than the number $N$ of observations, which is always tractable. For instance if one has a dataset of one billion rankings that each involve less than 5 items then the number of required operations is bounded by $5 \times 10^{11}$, which is still tractable.

From a theoretical point of view, the interesting aspect of the bound in Proposition 102 is that it does not depend directly on the number of items $n$. Only the term $\sum_{A\in\mathcal{A}} |A|!$ can

indeed depend on $n$ through the observation design $\mathcal{A}$, as explained in Subsection 3.1.6. More particularly, this term is exactly the bound on the number of parameters required to store the dataset $\mathcal{D}_N$ from Lemma 17. We can therefore say in a sense that the computation of the wavelet empirical estimator deals with the complexity of the data itself.

More generally, this can be considered as the great achievement of the MRA framework. As explained in Subsection 3.1.4, the analysis of incomplete rankings necessarily involves at some point the computation of the marginal $M_A q$ of a ranking model over $\mathfrak{S}_n$ on a subset of items $A \in \mathcal{P}([\![n]\!])$. If $q$ is represented as the vector of its values $(q(\sigma))_{\sigma \in \mathfrak{S}_n}$, the computation of $M_A q(\pi)$ for $\pi \in \Gamma(A)$ using Formula $(*)$ requires $n!/|A|!$ operations. Now, if $q$ is represented by its wavelet transform $\Psi q$, Theorem 57 tells us that $M_A q(\pi) = \phi_A \Psi q(\pi)$. The computation then has complexity bounded by $\binom{|A|}{2}$, by Proposition 84. This bound shows that the dependency in $n$ is an artifact of the theoretical framework of ranking models over $\mathfrak{S}_n$: when the ranking model is not represented as a function on $\mathfrak{S}_n$ but by its wavelet transform, this dependency vanishes.

## 5.2 Estimation of marginals

In this section we apply the MRA framework to the estimation of the marginals of a ranking model. It is certainly the problem where the MRA framework can be applied in the most straightforward manner. We thus use it to show the application of the MRA framework on a concrete problem but also to study some of its general properties.

### 5.2.1 Problem Statement and application of the MRA framework

As explained in Subsection 5.1.1, the accessible components of $p$ when observing incomplete rankings drawn from Process (3.3) are the wavelet projections $\Psi_B p$ for $B \in \bar{\mathcal{P}}(\mathcal{A})$ or equivalently the marginals $P_A = M_A p$ for $A \in \mathcal{A}$, where we recall that $\mathcal{A} = \mathrm{supp}(\nu)$ is considered to be known. The simplest problem we can consider is thus to estimate the marginals $(M_A p)_{A \in \mathcal{A}}$. For each $A \in \mathcal{A}$ we evaluate the quality of an estimator $\widehat{Q}_A$ of $P_A$ by the mean squared error (MSE) $\mathbb{E}[\|\widehat{Q}_A - P_A\|_A^2]$. We then define the error of a collection of estimators as the sum of the errors on each $A \in \mathcal{A}$ weighted by $\nu$.

**Definition 103** (Definition of the error)**.** The error of a collection of estimators $\widehat{Q} = (\widehat{Q}_A)_{A \in \mathcal{A}}$ with $\widehat{Q}_A \in L(\Gamma(A))$ for each $A \in \mathcal{A}$ is measured by

$$\mathcal{E}_N\left(\widehat{Q}\right) := \sum_{A \in \mathcal{A}} \nu(A) \mathbb{E}\left[\|\widehat{Q}_A - P_A\|_A^2\right].$$

*Remark* 104 (Possible negativity of the estimators). In the present setting, we don't impose to each estimator $\widehat{Q}_A$ to be a probability distribution over $\Gamma(A)$ for $A \in \mathcal{A}$. In particular for the MRA-based estimator, it can happen that a $\widehat{Q}_A^{MRA}$ takes negative values. This can be unfortunate in practice, if for instance one would like to use it to sample rankings or to compute conditional probabilities. This problem is actually classic in nonparametric statistics, usual methods to face it consist in approximating the estimator that takes negative values with the closest probability distribution in some sense (see for instance Huang et al., 2007). The drawback of such methods is of course that the final estimator is harder to control.

### 5.2.2 Application of the MRA framework

As announced, the application of the MRA framework to this setting is straightforward. The first step is to compute the wavelet empirical estimator $\widehat{\mathbf{X}} \in \mathbb{H}(\bar{\mathcal{P}}(\mathcal{A}))$ from the dataset $\mathcal{D}_N$ using

equation (5.1). Then one naturally defines an estimator of each $P_A$ using Theorem 57.

**Definition 105** (MRA-based estimator)**.** We define the MRA-based estimator $\widehat{Q}^{MRA} = (\widehat{Q}_A^{MRA})_{A\in\mathcal{A}}$ by

$$\widehat{Q}_A^{MRA} = \phi_A \widehat{\mathbf{X}} = \phi_A \sum_{B\in\overline{\mathcal{P}}(A)} \widehat{X}_B \qquad \text{for each } A \in \mathcal{A}.$$

Before illustrating the application on numerical experiments, we provide some theoretical guarantees about the error and the computational complexity of the MRA-based estimator.

**Theorem 106** (Statistical guarantees)**.** *The MRA-based estimator $(\widehat{Q}_A^{MRA})_{A\in\mathcal{A}}$ satisfies the two following properties.*

1. *It is "asymptotically unbiased", in the sense that:*

$$\lim_{N\to\infty} \mathbb{E}\left[\widehat{Q}_A^{MRA}\right] = P_A \qquad \text{for all } A \in \mathcal{A}.$$

2. *Its error decreases with a rate of order $O(1/N)$:*

$$\mathcal{E}_N\left(\widehat{Q}^{MRA}\right) \leq \frac{C_1}{N} + C_2\rho^{2N} \qquad \text{for all } N \geq 1,$$

*where $0 < \rho < 1$ is a constant that only depends on $\nu$ and $C_1$ and $C_2$ are positive constants that only depend on $p$ and $\nu$, given by $\rho = 1 - \min_{B\in\mathcal{P}(\mathcal{A})} \nu[\mathcal{Q}(B)]$,*

$$C_1 = 2\sum_{B\in\mathcal{P}(\mathcal{A})} \frac{\nu_\phi(B)}{\nu[\mathcal{Q}(B)]}\left(\|\Psi_B^2 p\|_{B,1} - \|\Psi_B p\|_{B,2}^2\right) \qquad \text{and} \qquad C_2 = \sum_{B\in\mathcal{P}(\mathcal{A})} \nu_\phi(B)\|\Psi_B p\|_{B,2}^2,$$

*where $\Psi_B^2 : L(\bar{\Gamma}_n) \to L(\Gamma(B))$ is the linear operator defined by $\Psi_B^2 F(\pi) = \sum_{\sigma\in\bar{\Gamma}_n} \alpha_B^2(\pi,\sigma_{|B})F(\sigma)$ for any $F \in L(\bar{\Gamma}_n)$, $\nu_\phi(B) := \sum_{A\in\mathcal{Q}(B)} 2^{|A|}\nu(A)/(|A|-|B|+1)!$ for any $B \in \mathcal{P}(\llbracket n\rrbracket)$ and $\nu[\mathcal{S}] := \sum_{A\in\mathcal{S}} \nu(A)$ for any collection of subsets $\mathcal{S} \subset \mathcal{P}(\llbracket n\rrbracket)$.*

Refer to the Appendix for the proof of Theorem 106. Property 1. is a natural consequence of Proposition 101. Property 2. relies on explicit calculations. If the constants $C_1$, $C_2$ and $\rho$ only depend on $p$ and $\nu$, and not directly on $n$, this is because the MRA representation enables to exploit only the part of information related to the observed dataset. We point out however that the more diffuse $\nu$ is, the more degrees of freedom the dataset has, and the bigger they are. The same interpretation applies to computational aspects.

**Proposition 107** (Computational guarantees)**.** *Let $K = \max_{A\in\mathcal{A}} |A|$ be the maximal size of an observed ranking.*

1. **Storage.** *The storage of the wavelet empirical estimator $\widehat{\mathbf{X}}$ requires a number of parameters upper bounded by $K!\, 2^K \min(N, |\mathcal{A}|)$.*

2. **Learning.** *The complexity of the computation of $\widehat{\mathbf{X}}$ is bounded by*

$$[e\, K! + (K+4)2^{K-1}] \min\left(N, \sum_{A\in\mathcal{A}} |A|!\right).$$

3. **Prediction.** *The computation of $\phi_A\widehat{\mathbf{X}}(\pi)$ for $A \in \mathcal{P}(\llbracket n\rrbracket)$ and $\pi \in \Gamma(A)$ needs less than $|A|(|A|-1)/2$ operations.*

*Proof.* By construction the MRA-based estimator can be stored as the collection of estimators $(\widehat{X}_B)_{B\in\mathcal{P}(\widehat{\mathcal{A}}_N)}$. The number of parameters to be stored is thus upper bounded by

$$\sum_{B\in\mathcal{P}(\widehat{\mathcal{A}}_N)} |B|! \leq K!\,|\mathcal{P}(\widehat{\mathcal{A}}_N)| \leq K! \sum_{A\in\widehat{\mathcal{A}}_N} 2^{|A|} \leq K!\,2^K|\widehat{\mathcal{A}}_N| \leq K!\,2^K \min(N, |\mathcal{A}|).$$

As for the other properties, Property 2 is already given by Proposition 102 and Property 3 is a direct consequence of Lemma 71. □

As in Theorem 106, the bounds in Proposition 107 do not depend directly on $n$, they only depend on the complexity of $\mathcal{A}$, the support of the observation design $\nu$. To give some more insights, the following example make the comparison with the empirical model (3.5) related to the approaches in Kondor and Barbosa (2010) and Sun et al. (2012).

*Example* 108. Consider the empirical model $\widehat{p}_N$ defined in Equation (3.5). It can be rewritten as

$$\widehat{p}_N = \sum_{A\in\widehat{\mathcal{A}}_N} \widehat{\nu}_N(A) \sum_{\pi\in\Gamma(A)} \widehat{P_A}(\pi)\mathbb{1}_{\mathfrak{S}_n(\pi)}.$$

Its most efficient storage is under the form of the collections of parameters $(\widehat{\nu}_N(A))_{A\in\widehat{\mathcal{A}}_N}$ and $(\widehat{P_A}(\pi))_{A\in\widehat{\mathcal{A}}_N,\,\pi\in\Gamma(A)}$, and the learning procedure is naturally in $O(N)$. But then, each computation of the marginal probability of a ranking $\pi' \in \Gamma_n$ involves the computation of all the inner products $\langle \mathbb{1}_{\mathfrak{S}_n(\pi')}, \mathbb{1}_{\mathfrak{S}_n(\pi)} \rangle$ for $\pi \in \bigsqcup_{A\in\widehat{\mathcal{A}}_N} \Gamma(A)$. This is at the root of the main computational limitation of the approaches introduced in Kondor and Barbosa (2010) and Sun et al. (2012).

### 5.2.3 Numerical Experiments

Here we examine the performance of the MRA-based estimator in numerical experiments and compare it with three others: the Plackett-Luce model (estimated by means of the MM algorithm from Hunter (2004)), the estimator from Sun et al. (2012), called SLK (we take the bandwidth of the kernel equal to $\binom{n}{2}+1$ to be sure that the smoothing is applied to the entire dataset), and the collection of naive empirical estimators $(\widehat{P_A})_{A\in\mathcal{A}}$.

Each experiment is characterized by a ranking model $p$, a probability distribution $\nu$ and a number of observations $N$. We consider two theoretical ranking models, namely a Plackett-Luce model defined with parameter vector $\mathbf{w} = (w_1, \ldots, w_n)$ drawn uniformly at random on the simplex $\{\mathbf{x} \in [0,1]^n \mid \sum_{i=1}^n x_i = 1\}$ and a Mallows model defined for $\sigma \in \mathfrak{S}_n$ by $p(\sigma) \propto e^{-d_{KT}(\sigma_0,\sigma)}$ where $\sigma_0 = 12\ldots n$, and one empirical model, namely the distribution of the 5738 votes in the APA dataset (from Diaconis, 1989) that we consider as a ground truth ranking model. In all the experiments, $n = 5$. For each ranking model, we examine the four different settings where $\nu$ is the uniform probability distribution on $\{A \subset [\![5]\!] \mid 2 \leq |A| \leq k\}$ for $k = 2, 3, 4, 5$, and let the size $N$ of the drawn dataset $\mathcal{D}_N$ vary between 500 and 5000. We then evaluate the performance of an estimator $\widehat{Q}$ constructed from $\mathcal{D}_N$ through a Monte-Carlo estimate of $\mathcal{E}_N(\widehat{Q})$ averaged from 100 drawings of $\mathcal{D}_N$.

Figure 5.1 depicts the experimental results. As explained in Subsection 3.1.5, the SLK ranking model applies a strong smoothing which leads to a very small variance but an important bias when $p$ differs from the uniform distribution on $\mathfrak{S}_n$. This is why it converges rapidly and its performance is almost constant through the experiments for $N \geq 500$. The Plackett-Luce model relies on a structural assumption and is thus naturally biased when $p$ is not a Plackett-Luce model. This explains why it does not perform best in the latter case. The MRA-based estimator and the naive empirical estimators are both asymptotically unbiased whatever the underlying

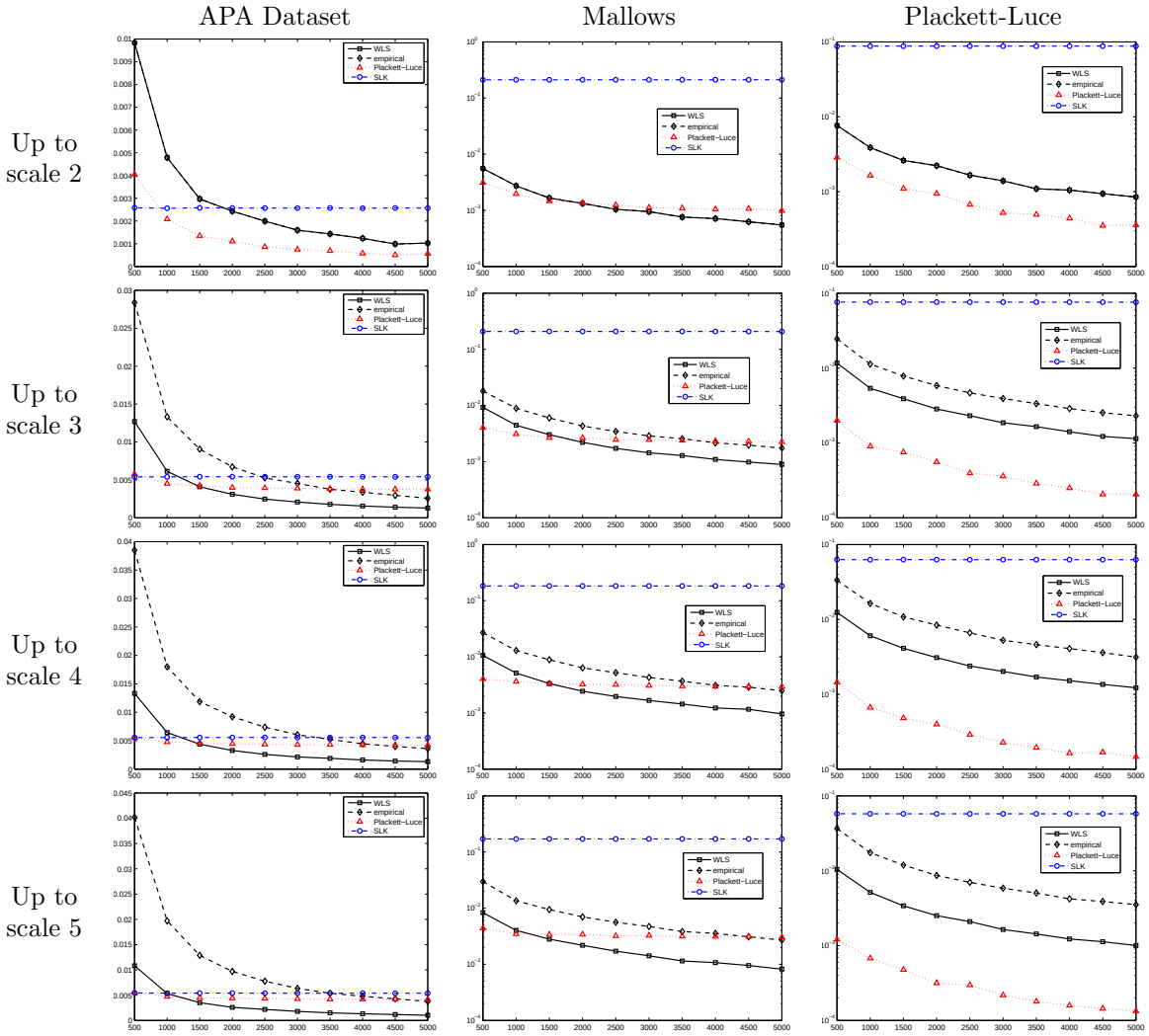



Figure 5.1: Evolution of the performance $\mathcal{E}_N\left(\widehat{Q}\right)$ with $N$ for each estimator: MRA-based in black squares (written WLS), Naive empirical estimator in black diamonds, Plackett-Luce in red triangles and SLK in blue circles, with different underlying ranking models: APA dataset (first column), Mallows (second column), Plackett-Luce (third column) and with probability $\nu$ uniform on $\{A \subset [\![5]\!] \mid 2 \leq |A| \leq k\}$ with $k = 2, 3, 4, 5$ (from top to bottom). For the Mallows and Plackett-Luce models, the results are represented on a logarithmic scale.

ranking model $p$ and have similar behaviors, except that the MRA-based estimator has reduced variance and thus converges faster. Globally, the MRA-based estimator quickly outperforms its competitors when $N$ grows.

## 5.3 Ranking prediction on a subset

In this section we consider the problem of ranking prediction: for any subset of elements $A \in \mathcal{P}(\llbracket n \rrbracket)$ we want to predict the "best" ranking $\pi \in \Gamma(A)$ on $A$. In an e-commerce setting, this ranking could correspond to the order in which the items of $A$ should be presented to the users of a homogeneous population to best fit their preferences.

### 5.3.1 Problem statement

This problem is naturally stated in a supervised learning framework: the input space is $\mathcal{P}(\llbracket n \rrbracket)$, the output space $\Gamma_n$, and a classifier is a mapping $\mathcal{P}(\llbracket n \rrbracket) \to \Gamma_n$, with the particularity that it must map a subset of elements $A$ to a ranking on $A$. An equivalent point of view is to see a classifier as a collection $\widehat{\pi} = (\widehat{\pi}_A)_{A \in \mathcal{P}(\llbracket n \rrbracket)}$ where $\widehat{\pi}_A \in \Gamma(A)$ is a ranking on $A$ for each $A \in \mathcal{P}(\llbracket n \rrbracket)$.[1] We thus define the hypothesis space $\mathcal{H}$ as

$$\mathcal{H} = \prod_{A \in \mathcal{P}(\llbracket n \rrbracket)} \Gamma(A).$$

We assume to observe incomplete rankings as samples of a random couple $(\mathbf{A}, \Pi)$ drawn according to Process (3.3) from a ranking model $p$ and a probability distribution $\nu$ over $\mathcal{P}(\llbracket n \rrbracket)$. The ranking model $p$ is unknown and characterizes the preferences of the statistical population. The probability distribution $\nu$ is known but imposed. In an e-commerce setting, this means that users express their preferences as incomplete rankings, over subsets of items that they choose or that stem from an independent context (defined by navigation filters for instance).

The accuracy of a classifier is thus evaluated against incomplete rankings that represent the ground truth with a loss function of the form:

$$d : \bigsqcup_{A \in \mathcal{P}(\llbracket n \rrbracket)} \Gamma(A) \times \Gamma(A) \to \mathbb{R}_+.$$

We assume that $d$ is such that for each $A \in \mathcal{P}(\llbracket n \rrbracket)$, the restriction $d_{|\Gamma(A)^2}$ of $d$ to $\Gamma(A)^2$ is a distance on $\Gamma(A)$ (refer to Subsection 2.4.4 for examples of distances). In order to ensure a consistent evaluation on varying subsets, of potentially different sizes, we assume that $d$ satisfies three conditions:

1. It is invariant under relabeling of the elements of $\llbracket n \rrbracket$: for any $A \in \mathcal{P}(\llbracket n \rrbracket)$, $\pi, \pi' \in \Gamma(A)$ and $\tau \in \mathfrak{S}_n$, $d(\tau(\pi), \tau(\pi')) = d(\pi, \pi')$.
2. It is normalized so that $\max_{(\pi,\pi') \in \Gamma(A)^2} d(\pi, \pi') = 1$ for all $A \in \mathcal{P}(\llbracket n \rrbracket)$.[2]
3. All the restrictions $d_{|\Gamma(A)^2}$ "represent the same distance" for $A \in \mathcal{P}(\llbracket n \rrbracket)$, in the sense that if $d_{|\Gamma(A)^2}$ is for example the Kendall's tau distance on $\Gamma(A)$ then $d_{|\Gamma(B)^2}$ is the Kendall's tau distance on $\Gamma(B)$ for all $B \in \mathcal{P}(\llbracket n \rrbracket)$.

[1] We acknowledge that the notation $\widehat{\cdot}$ can be misleading because it is used in the rest of this thesis to designate stochastic objects. In most cases though, $\widehat{\pi}$ will designate a classifier constructed from a dataset and therefore the notation will be consistent.

[2] This choice is arbitrary but it is not our purpose here to analyze its impact.

The theoretical risk for the problem of incomplete rankings prediction is then defined, for a classifier $\widehat{\pi}$, as the expectation

$$\mathcal{R}(\widehat{\pi}) = \mathbb{E}_{\mathbf{A},\Pi}[d\left(\widehat{\pi}_{\mathbf{A}}, \Pi\right)] = \sum_{A \in \mathcal{P}(\llbracket n \rrbracket)} \nu(A) \sum_{\pi \in \Gamma(A)} d\left(\widehat{\pi}_A, \pi\right) P_A(\pi). \tag{5.2}$$

As in the classic supervised learning framework, the goal of incomplete ranking prediction is to find a classifier that minimizes this theoretical risk. Of course, as the ranking model $p$ is not known, we are led to consider the empirical version of the risk, defined for a dataset $\mathcal{D}_N = ((\mathbf{A}_1, \Pi^{(1)}), \ldots, (\mathbf{A}_N, \Pi^{(N)}))$ of $N$ IID samples of $(\mathbf{A}, \Pi)$ by

$$\widehat{\mathcal{R}}_N(\widehat{\pi}) = \frac{1}{N} \sum_{i=1}^{N} d\left(\widehat{\pi}_{A_i}, \Pi^{(i)}\right). \tag{5.3}$$

We point out that because rankings can only be observed on subsets $A \in \mathcal{A}$ a classifier $\widehat{\pi} \in \mathcal{H}$ is only evaluated through its values $\widehat{\pi}_A$ for $A \in \mathcal{A}$.

### 5.3.2 General analysis and application of the MRA framework

Here we provide more insights about the challenges at stake in the problem of ranking prediction on a subset and detail the application of the MRA framework.

**Optimality and empirical risk minimization.** The first result we give exhibits an optimal classifier for the stated problem. Its proof is straightforward and left to the reader.

**Proposition 109** (Optimal classifier)**.** *Let $\widehat{\pi}^*$ be a classifier such that $\widehat{\pi}^*_A$ is a solution of the minimization problem*

$$\min_{\pi \in \Gamma(A)} \sum_{\pi' \in \Gamma(A)} d(\pi, \pi') P_A(\pi') \tag{5.4}$$

*for all $A \in \mathcal{A}$. Then, the classifier $\widehat{\pi}$ has minimum risk (5.2).*

We point out that problem (5.4) is a rank aggregation problem, it corresponds for each $A \in \mathcal{A}$ to (2.2) with rankings in $\Gamma(A)$ and distance $d$. In particular an optimal classifier $\widehat{\pi}^*$ always exists: the set $\Gamma(A)$ is of finite cardinality and, thus, there always exists a solution to the minimization problem (5.4). It is however not necessarily unique.

*Remark* 110. The present setting resembles the one of multi-class classification, but the definition (5.4) of the optimal classifier depends here on the loss function. For the $0-1$ loss function defined by $d(\pi, \pi') = \mathbb{I}\{\pi \neq \pi'\}$, equation (5.4) becomes $\min_{\pi_0 \in \Gamma(A)} \sum_{\pi \neq \pi_0} P_A(\pi) = \min_{\pi_0 \in \Gamma(A)} (1 - P_A(\pi_0))$, and the optimal classifier $\widehat{\pi}^*$ is defined by $\widehat{\pi}^*_A = \operatorname{argmax}_{\pi \in \Gamma(A)} P_A(\pi)$, that is to say $\widehat{\pi}^*$ is equal to the optimal Bayes classifier for the corresponding multi-class classification problem. This is not true in general.

As in classic supervised learning, one cannot of course compute an optimal classifier $\widehat{\pi}^*$ in practice because the $P_A$'s are not known. The usual approach to define a classifier that approximates $\widehat{\pi}^*$ is through empirical risk minimization. This principle can be applied directly for the general hypothesis space of all possible classifiers $\mathcal{H} = \prod_{A \in \mathcal{P}(\llbracket n \rrbracket)} \Gamma(A)$ in the present setting, since it is finite. It characterizes however a classifier only for the subsets of elements $A$ that were observed: $A \in \widehat{\mathcal{A}}_N$. For any $A \in \mathcal{P}(\llbracket n \rrbracket)$ we define $\widehat{I}_A = \{1 \leq i \leq N \mid \mathbf{A}_i = A\}$ the set of indexes that correspond to the observation of $A$, so that $A \in \widehat{\mathcal{A}}_N$ is equivalent to $\widehat{I}_A \neq \emptyset$.

**Definition 111** (Empirical risk minimization over $\mathcal{H}$)**.** The solutions to the empirical risk minimization problem $\min_{\widehat{\pi}\in\mathcal{H}} \widehat{\mathcal{R}}_N(\widehat{\pi})$ are the classifiers $\widehat{\pi}^{ERM}$ that satisfy for each $A \in \widehat{\mathcal{A}}_N$

$$\widehat{\pi}_A^{ERM} = \min_{\pi\in\Gamma(A)} \sum_{i\in\widehat{I}_A} d\left(\pi, \Pi^{(i)}\right). \tag{5.5}$$

While the rankings of an optimal classifier are solutions of an inference rank aggregation problem of type (2.2), each ranking $\widehat{\pi}_A^{ERM}$ is a solution to (5.5), which corresponds to (2.1) for rankings in $\Gamma(A)$ and distance $d$. Each ranking $\widehat{\pi}_A^{ERM}$ is thus obtained from "local" aggregation on $A$ of the rankings $(\mathbf{A}_i, \Pi^{(i)})_{i\in\widehat{I}_A}$. Now, as the dataset $\mathcal{D}_N$ can be partitioned into a collection of observations on a same subset $A \in \widehat{\mathcal{A}}_N$: $\mathcal{D}_N = \bigsqcup_{A\in\widehat{\mathcal{A}}_N}(\mathbf{A}_i, \Pi^{(i)})_{i\in\widehat{I}_A}$, the $\widehat{\pi}_A^{ERM}$ are all independent. Therefore they do not consolidate information between observations on different subsets and can be highly variable. For instance if a subset $A \in \mathcal{P}(\llbracket n \rrbracket)$ only appears once in $\mathcal{D}_N$ then the ERM classifier will predict on $A$ the sole ranking it has observed on $A$. It would not try to infer what could be potential observations on $A$ from other drawings of $\mathcal{D}_N$ from observations on other subsets. This motivates the design of other classifiers.

*Remark* 112 (Comparison with multi-task/transfer learning)*.* Incomplete ranking prediction can be seen as a multi-task learning problem, where each task corresponds to the prediction on a fixed subset $A \in \mathcal{P}(\llbracket n \rrbracket)$. These tasks are related here by the consistency assumption (∗). As in multi-task learning, the efficiency is increased when all the tasks are learned jointly.

**Rank aggregation.** A natural approach to consolidate information between observation on different subsets is to perform global rank aggregation over $\llbracket n \rrbracket$ and use the obtained full ranking to induce rankings on any subset. This approach is formalized in the following definition.

**Definition 113** (Global aggregation-based classifier)**.** A global aggregation-based classier is a classifier $\widehat{\pi}^{agg} \in \mathcal{H}$ defined from a full ranking $\sigma \in \mathfrak{S}_n$ by

$$\widehat{\pi}_A^{agg} = \sigma_{|A} \qquad \text{for each } A \in \mathcal{P}(\llbracket n \rrbracket).$$

The most natural way to construct the full ranking $\sigma$ from the dataset $\mathcal{D}_N$ would certainly be to compute it such that the associated classifier minimizes the empirical risk (5.3). It would thus be a solution of the minimization problem

$$\min_{\sigma'\in\mathfrak{S}_n} \sum_{i=1}^{N} d\left(\sigma'_{|\mathbf{A}_i}, \Pi^{(i)}\right). \tag{5.6}$$

Problem (5.6) can be too costly to solve exactly in practice but many approaches exist to compute an approximate solution, either via a parametric model (see Subsection 2.5.1) or with a specific procedure for rank aggregation from incomplete rankings (see 2.3).

The strength of the global aggregation-based classifiers paradigm is that it provides a simple approach to consolidate information from observations on different subsets. It imposes however the strong constraint on the obtained classifier $\widehat{\pi}^{agg}$ that all the rankings $\widehat{\pi}_A^{agg}$ must be consistent: for any pair of elements $\{a, b\} \subset \llbracket n \rrbracket$ and subsets $A, A' \subset \llbracket n \rrbracket$ that contain $\{a, b\}$, the rankings $\widehat{\pi}_A^{agg}$ and $\widehat{\pi}_{A'}^{agg}$ must rank $a$ and $b$ in the same order. Yet, while the $P_A$'s must satisfy the consistency assumption (∗), there is no reason that an optimal classifier $\widehat{\pi}^*$ should satisfy this constraint. This is illustrated by the following example.

*Example* 114. Let $n = 4$ and $p$ be the ranking model defined by

$$p = \frac{1}{4}\left[\delta_{1234} + \delta_{4123} + \delta_{3412} + \delta_{2341}\right].$$

Notice that on a pair of elements $\{a, b\} \subset [\![n]\!]$, $\widehat{\pi}_{\{a,b\}}$ is the ranking $\pi \in \{ab, ba\}$ that has maximal probability $P_{\{a,b\}}(\pi)$, whatever the distance $d$. One thus has in the present example.

| Pair $\{a, b\}$ | $P_{\{a,b\}}(ab)$ | $P_{\{a,b\}}(ba)$ | $\widehat{\pi}^*_{\{a,b\}}$ |
|---|---|---|---|
| $\{1, 2\}$ | 3/4 | 1/4 | 12 |
| $\{1, 3\}$ | 1/2 | 1/2 | $\{13, 31\}$ |
| $\{1, 4\}$ | 1/4 | 3/4 | 41 |
| $\{2, 3\}$ | 3/4 | 1/4 | 23 |
| $\{2, 4\}$ | 1/2 | 1/2 | $\{24, 42\}$ |
| $\{3, 4\}$ | 3/4 | 1/4 | 34 |

There is no full ranking $\sigma \in \mathfrak{S}_4$ such that

$$\sigma_{\{1,2\}} = 12, \quad \sigma_{\{2,3\}} = 23, \quad \sigma_{\{3,4\}} = 34, \quad \text{and} \quad \sigma_{\{1,4\}} = 41.$$

For a general $n$, the ranking model $p = (1/n)[\delta_{12\dots n} + \delta_{n1\dots(n-1)} + \cdots + \delta_{23\dots1}]$ satisfies the same property.

Example 114 can seem a little artificial but numerical experiments show that this phenomenon happens on empirical datasets (see Subsection 5.3.3).

**Plug-in paradigm.** We now introduce a general approach to construct a classifier from the dataset without imposing the constraint of global aggregation. It is based on the following intuition: as an optimal classifier is obtained as local consensuses for the true marginals $P_A$, computing local consensuses for good estimators of the $P_A$'s should provide a good classifier. The approach we propose thus consists in first constructing estimators for the $P_A$'s then computing the local consensuses. We first introduce the following definition.

**Definition 115** (Generalized consensus)**.** Let $A \in \mathcal{P}([\![n]\!])$ and $d$ be a metric on $\Gamma(A)$. A ranking $\pi^* \in \Gamma(A)$ is a consensus for a function $F \in L(\Gamma(A))$ if it satisfies

$$\pi^* = \operatorname*{argmin}_{\pi \in \Gamma(A)} \sum_{\pi' \in \Gamma(A)} d(\pi, \pi') F(\pi').$$

We denote by $\mathcal{C}_d(F) \subset \Gamma(A)$ the set of consensus rankings for $F \in L(\Gamma(A))$ with respect to $d$.

Definition 115 corresponds to (2.2) for rankings on a subset $A \in \mathcal{P}([\![n]\!])$ and distance $d$, generalized to functions (not just probability distributions). Though it is not common to consider this concept of generalized consensuses in the literature, it was previously introduced (for instance in Saari, 2000). It is exploited to give more insights about rank aggregation in Subsection 6.3.3 but here we mostly use it to introduce the following definition.

**Definition 116** (Plug-in paradigm)**.** For a family of functions $Q = (Q_A)_{A \in \mathcal{P}([\![n]\!])}$ with $Q_A \in L(\Gamma(A))$ for each $A \in \mathcal{P}([\![n]\!])$, an associated plug-in classifier $\widehat{\pi}(Q) \in \mathcal{H}$ is a classifier such that $\widehat{\pi}_A(Q) \in \mathcal{C}_d(Q_A)$ for all $A \in \mathcal{P}([\![n]\!])$.

As explained previously, optimal classifiers and minimizers of the empirical risk can be seen as plug-in classifiers. It is also the case for global aggregation-based classifiers, because $\mathcal{C}_d(\delta_\pi) = \{\pi\}$ for any subset $A \in \mathcal{P}([\![n]\!])$ ranking $\pi \in \Gamma(A)$ and distance $d$. One thus has, for an optimal classifier $\widehat{\pi}^*$, a minimizer of the empirical risk $\widehat{\pi}^{ERM}$ and a global aggregation-based classifier $\widehat{\pi}^{agg}$ constructed from the full ranking $\sigma \in \mathfrak{S}_n$

$$\widehat{\pi}^* = \widehat{\pi}\left((P_A)_{A \in \mathcal{P}([\![n]\!])}\right), \qquad \widehat{\pi}^{ERM} = \widehat{\pi}\left((\widehat{P_A})_{A \in \mathcal{P}([\![n]\!])}\right) \qquad \text{and} \qquad \widehat{\pi}^{agg} = \widehat{\pi}\left((\delta_{\sigma_{|A}})_{A \in \mathcal{P}([\![n]\!])}\right).$$

The plug-in paradigm we develop here is analogous to the classic plug-in paradigm in supervised learning, (see Audibert and Tsybakov, 2007; Clémençon and Robbiano, 2011). It relies on the fact that if $(Q_A)_{A\in\mathcal{P}(\llbracket n\rrbracket)}$ is close in a certain sense to $(P_A)_{A\in\mathcal{P}(\llbracket n\rrbracket)}$ then the risk of $\widehat{\pi}((Q_A)_{A\in\mathcal{P}(\llbracket n\rrbracket)})$ should be close to the risk of $\widehat{\pi}((P_A)_{A\in\mathcal{P}(\llbracket n\rrbracket)})$, that is to say the minimum risk. This is guaranteed by the following proposition. For $A \in \mathcal{P}(\llbracket n\rrbracket)$, we define the $|A|! \times |A|!$ matrix $D_A$ by $D_A(\pi, \pi') = d(\pi, \pi')$ for $\pi, \pi' \in \Gamma(A)$ and denote by $\|\cdot\|_{A,\infty}$ the infinity norm on $L(\Gamma(A))$, defined by $\|F\|_{A\infty} = \max_{\pi\in\Gamma(A)} |F(\pi)|$ for any $F \in L(\Gamma(A))$.

**Proposition 117.** *Let $\widehat{\pi}^*$ be an optimal classifier. Then for any family of functions $Q = (Q_A)_{A\in\mathcal{P}(\llbracket n\rrbracket)}$,*

$$\mathcal{R}(\widehat{\pi}(Q)) - \mathcal{R}(\widehat{\pi}^*) \;\leq\; 2\sum_{A\in\mathcal{A}} \nu(A)\|D_A\left(Q_A - P_A\right)\|_{A,\infty}$$

*Proof.* By definition for any classifier $\widehat{\pi} = (\widehat{\pi}_A)_{A\in\mathcal{P}(\llbracket n\rrbracket)} \in \mathcal{H}$

$$\mathcal{R}(\widehat{\pi}) = \sum_{A\in\mathcal{A}} \nu(A) \sum_{\pi\in\Gamma(A)} d(\widehat{\pi}_A, \pi) P_A(\pi) = \sum_{A\in\mathcal{A}} \nu(A) D_A P_A(\widehat{\pi}_A).$$

Let $A \in \mathcal{P}(\llbracket n\rrbracket)$. For $\pi \in \mathcal{C}_d(Q_A)$ and $\pi^* \in \mathcal{C}_d(P_A)$,

$$\begin{aligned}
D_A P_A(\pi) - D_A P_A(\pi^*) &= D_A(P_A - Q_A)(\pi) + D_A Q_A(\pi) - D_A P_A(\pi^*) \\
&= D_A(P_A - Q_A)(\pi) + \min_{\pi'\in\Gamma(A)} D_A Q_A(\pi') - \min_{\pi'\in\Gamma(A)} D_A P_A(\pi') \\
&\leq \max_{\pi'\in\Gamma(A)} D_A(P_A - Q_A)(\pi) + \max_{\pi'\in\Gamma(A)} (Q_A - P_A)(\pi') \\
&\leq 2\|D_A\left(Q_A - P_A\right)\|_{A,\infty}.
\end{aligned}$$

Summing over the subsets $A$ gives the desired result. □

**MRA-based classifier.** We now define the MRA-based classifier.

**Definition 118** (MRA-based classifier)**.** The MRA-based classifier is the plug-in classifier associated with the MRA-based estimator $\widehat{Q}^{MRA}$ from Definition 105

$$\widehat{\pi}^{MRA} := \widehat{\pi}\left(\widehat{Q}^{MRA}\right).$$

The statistical guarantees about the MRA-based estimator $\widehat{Q}^{MRA}$ transfer into statistical guarantees about the MRA-based classifier.

**Theorem 119** (Theoretical guarantees)**.** *Let $\widehat{\pi}^*$ be an optimal classifier. One then has*

$$\mathbb{E}\left[\left(\mathcal{R}\left(\widehat{\pi}^{MRA}\right) - \mathcal{R}\left(\widehat{\pi}^*\right)\right)^2\right] \leq 4C^2\left(\frac{C_1}{N} + C_2\rho^N\right),$$

*where $C = \sum_{\pi\in\Gamma(\llbracket K\rrbracket)} d(1\dots k, \pi)$ with $K = \max_{A\in\mathcal{A}} |A|$ and $C_1$, $C_2$ and $\rho$ are the constants from Theorem 106.*

*Proof.* By Proposition 117 one has

$$\mathbb{E}\left[\left(\mathcal{R}\left(\widehat{\pi}^{MRA}\right) - \mathcal{R}\left(\widehat{\pi}^*\right)\right)^2\right] \leq \mathbb{E}\left[\left(2\sum_{A\in\mathcal{A}} \nu(A)\left\|D_A\left(\widehat{Q}_A^{MRA} - P_A\right)\right\|_{A,\infty}\right)^2\right].$$

For $A \in \mathcal{A}$, $\|D_A\left(\widehat{Q}_A^{MRA} - P_A\right)\|_{A,\infty} \leq \|D_A\|_\infty \|\widehat{Q}_A^{MRA} - P_A\|_{A,\infty}$, where $\|D_A\|_\infty$ is the matrix norm of $D_A$ induced by $\|\cdot\|_{A,\infty}$. It is well known that $\|D_A\|_\infty = \max_{\pi\in\Gamma(A)} \sum_{\pi'\in\Gamma(A)} |d(\pi,\pi')|$. Since $d$ is invariant under relabeling, one has $\|D_A\|_\infty = \sum_{\pi\in\Gamma([\![|A|]\!])} d(1\dots|A|,\pi) \leq C$. Injecting this result and using the Cauchy-Schwarz inequality one obtains

$$\begin{aligned}\mathbb{E}\left[\left(\mathcal{R}\left(\widehat{\pi}^{MRA}\right) - \mathcal{R}\left(\widehat{\pi}^*\right)\right)^2\right] &\leq 4C^2\mathbb{E}\left[\left(\sum_{A\in\mathcal{A}}\nu(A)\right)\left(\sum_{A\in\mathcal{A}}\nu(A)\left\|\widehat{Q}_A^{MRA} - P_A\right\|_{A,\infty}^2\right)\right] \\ &\leq 4C^2\mathbb{E}\left[\sum_{A\in\mathcal{A}}\nu(A)\left\|\widehat{Q}_A^{MRA} - P_A\right\|_A^2\right],\end{aligned}$$

where the last inequality uses the facts that $\sum_{A\in\mathcal{A}}\nu(A) = 1$ and $\|F\|_{A,\infty} \leq \|F\|_A$ for all $F \in L(\Gamma(A))$ with $A \in \mathcal{P}([\![n]\!])$. The proof is concluded using Theorem 106. □

The computational guarantees of the MRA-based estimator from Proposition 107 directly apply to the MRA-based classifier. One must simply add to the prediction on a subset $A \in \mathcal{P}([\![n]\!])$ the cost to compute $\widehat{\pi}_A$ from $\widehat{Q}_A$, that is to say to compute a consensus for $\widehat{Q}_A$. As predictions are made on small subsets, this can be made with a brute-force search with complexity $|A|!$.

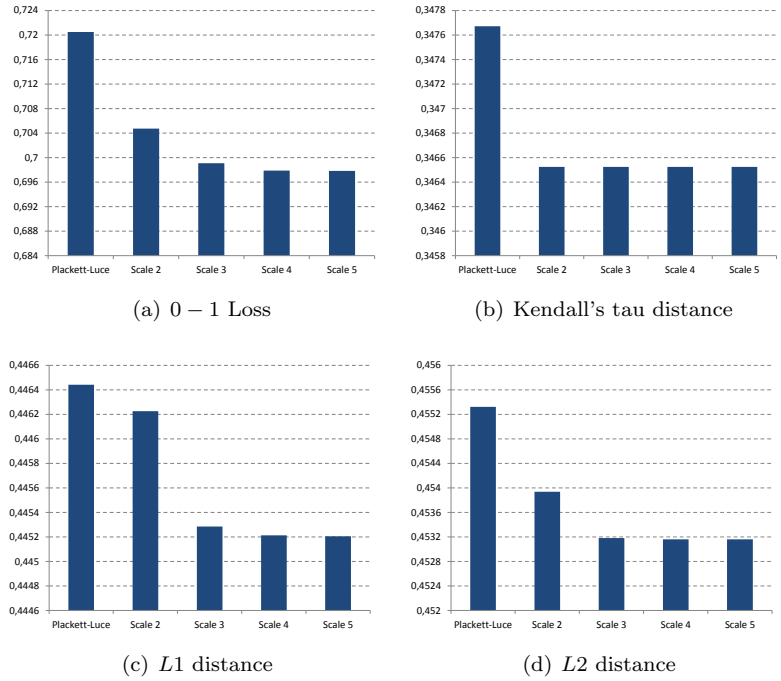
### 5.3.3 Numerical experiments

We present the results of numerical experiments conducted on data generated from two real datasets, the SUSHI dataset ($n = 10$) and the NETFLIX dataset ($n = 17,770$). In both cases, we generate from raw data incomplete rankings of size 2 to 5 (the maximum size 5 is a consequence of the rating scale in the NETFLIX dataset).

**Evaluation setting.** The predictions are evaluated through four different distances: the $0-1$ loss, Kendall's tau, Spearman's footrule and Spearman's rho (see Subsection 2.4.4). All these distances are invariant under relabeling of the items, and can thus evaluate the accuracy of the predictions on different subsets of items of the same size in a consistent manner. In order to be fully consistent when dealing with subsets of different sizes, we use their normalized versions defined for $\pi, \pi' \in \Gamma(A)$ with $A \in \mathcal{P}([\![n]\!])$ by $\overline{d}(\pi,\pi') = d(\pi,\pi')/d_{|A|}$, where $d$ is one of the four distances and $d_k = \max_{\pi,\pi'\in\Gamma(\{1,\dots,k\})} d(\pi,\pi')$ for $k \in \{2,\dots,5\}$.

A classifier is evaluated by its empirical risk on a test dataset with respect to a loss function. As a baseline, we compute the expectation and standard deviation of the risk of the uniformly random classifier that predicts a ranking in $\Gamma(A)$ drawn uniformly at random, for any subset of items $A \in \mathcal{P}([\![n]\!])$. We compare the performance of the MRA-based classifier with the classifier based on global aggregation by the Plackett-Luce model (fitted by means of the MM algorithm from Hunter (2004)), which happens to be the same as the plug-in classifier based on the Plackett-Luce model for all considered distances (this fact surely has a theoretical explanation but we are not aware of it, it may constitute an interesting direction for future research).

**Experiments based on the Sushi dataset.** The SUSHI dataset from Kamishima (2003) is composed of 5000 full rankings on 10 sushi varieties. By drawing from each full ranking 210 incomplete rankings of size 2 to 5 uniformly at random, we generate a global dataset of $1,260,000$ incomplete rankings, for which we keep 80% as a training set and 20% as a test set. We evaluate the estimator $\tilde{p}_N$ and also its truncated versions to scales $k = 2$, 3, or 4, where the $\widehat{X}_B$'s are put equal to 0 for $|B| > k$. The results are shown on Figure 5.2. They represent the empirical risks on the test set for the plug-in classifiers of the five probabilistic models for the




(a) $0-1$ Loss (b) Kendall's tau distance

(c) $L1$ distance (d) $L2$ distance

Figure 5.2: Empirical risk of the plug-in predictors on the SUSHI dataset

| | $0-1$ loss | Kendall's tau | Spearman's footrule | Spearman's rho |
|---|---|---|---|---|
| Expectation | 0.8208 | 0.5000 | 0.6145 | 0.6156 |
| Standard deviation | $6.6 \times 10^{-4}$ | $6.7 \times 10^{-4}$ | $7.1 \times 10^{-4}$ | $6.8 \times 10^{-4}$ |

Table 5.1: Expectation and standard deviation of the uniformly random classifier for the SUSHI dataset

four loss functions. All plug-in classifiers are computed exactly. As a baseline, the expectation and standard deviation of the uniformly random predictor are given in Table 5.1.

In each case, the risk of the worst model is lower than that of the uniformly random predictor by hundreds of standard deviations. This is surely explained by the fact that all statistical models manage to leverage information from historical data to make better predictions than random, and the amount of the difference is due to the large size of the test set $(252,000)$. For all four distances, all the multiresolution classifiers outperform the one based on the Plackett-Luce model. This demonstrates the pertinence and accuracy of our approach. An interesting observation is that for each loss function, the risk of the truncated multiresolution-based predictor decreases with the scale. This means that each scale contains a specific part of information that is useful to make better predictions. It shows in particular that reducing the observations to pairwise comparisons inherently degrades the available information, and proves the interest to exploit higher order information.

**Experiments based on the Netflix dataset.** The NETFLIX dataset was issued for the *Netflix Prize*. The training set contains $100,480,507$ ratings given by $480,189$ users to $17,770$ movies. Each rating is an integer between 1 and 5. We use the training set to generate incomplete

rankings, on the following simple paradigm: if a user gave respectively the ratings $r_a$ and $r_b$ to movies $a$ and $b$ with $r_a > r_b$ then it means that she prefers movie $a$ to movie $b$. More generally if she gave the ratings $r_1 > \cdots > r_k$ to the movies $a_1, \ldots, a_k$, her preference over the subset of movies $a_1, \ldots, a_k$ is given by the ranking $a_1 \ldots a_k$. As the grades are on a scale from 1 to 5, the obtained incomplete rankings are of maximum size 5. For each user, we consider the list of the ratings she gave, draw uniformly at random subsets of movies, and generate the corresponding incomplete rankings. We keep the first 80% ratings for training and the last 20% for test. We then aggregate the data to obtain a training set and a test set of respectively $153,703,541$ and $38,665,610$ incomplete rankings.

For computational reasons, we only tested the predictive rule consisting in choosing the ranking with higher probability, for both the MRA-based classifier and the one based on the Plackett-Luce model. We nevertheless evaluated their performance through the four loss functions considered in the previous section. The results, as well as the expectation and standard deviation of the uniformly random classifier, are presented in table 5.2.

| | $0-1$ loss | Kendall's tau | Spearman's footrule | Spearman's rho |
|---|---|---|---|---|
| Expectation | 0.7388 | 0.5000 | 0.6059 | 0.5867 |
| Standard deviation | $6.5 \times 10^{-5}$ | $6.1 \times 10^{-5}$ | $6.6 \times 10^{-5}$ | $6.3 \times 10^{-5}$ |
| Plackett-Luce | 0.5579 | 0.3598 | 0.3934 | 0.3865 |
| MRA | 0.6042 | 0.3938 | 0.4425 | 0.4328 |

Table 5.2: Results for the NETFLIX dataset

Again, both statistical models outperform by far the random classifier. Contrary to the SUSHI dataset, the Plackett-Luce model performs better than the MRA-based estimator for all loss functions. This is surely due to the fact that the classifier based on the global aggregation from the Plackett-Luce model captures global effects on the full set $[\![n]\!]$, namely the average rank of a movie in any incomplete ranking. It is indeed highlighted in Koren (2009) that the tendencies of some movies to receive higher ratings than others captures much of the information in the NETFLIX dataset. On the contrary the MRA-based classifier is best fitted to capture pure relative preferences effects. Its application in such large-scale settings should thus be made with a regularization procedure (see Section 7.1 for some propositions). In any case, this experiment demonstrates the good scalability of the MRA framework.

# Chapter 6

# Connections and other interpretations

Perhaps surprisingly, the MRA representation draws connections between many mathematical constructions related to permutations and rankings. We detail them in this chapter, together with the insights they provide. First, the connection with Fourier analysis is made in Section 6.1. Then in Section 6.2 we establish many connections with diverse constructions (related to card shuffling, generalized Kendall's tau distances) through the study of the alternative embedding operator from Subsection 4.1.3. At last we focus on the component related to pairwise information in Section 6.3 and establish new results about Kemeny rank aggregation.

## Contents



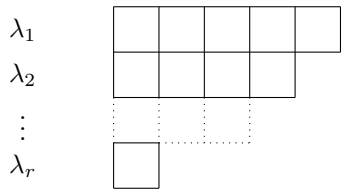
## 6.1 Connection with Fourier analysis

We begin with the connection between the MRA representation and Fourier analysis (refer to Subsection 3.2.5 for some background about Fourier analysis on $\mathfrak{S}_n$).

### 6.1.1 Background on Young tableaux

We recall that in the Fourier decomposition of $L(\mathfrak{S}_n)$ (see Proposition 34), each irreducible representation $S^\lambda$ appears with multiplicity $d_\lambda$ for $\lambda \vdash n$. It happens that $d_\lambda$ is also the number

of different *standard Young tableaux* of shape $\lambda$. One can therefore enumerate the copies of a same irreducible representation $S^\lambda$ with such objects, which will be useful to establish the connection with the MRA. Let us introduce some definitions to be more specific. A Young diagram (or a Ferrer's diagram) of size $n$ is a collection of boxes of the form




where if $\lambda_i$ denotes the number of boxes in row $i$, then $\lambda = (\lambda_1, \dots, \lambda_r)$, called the shape of the Young diagram, must be a partition of $n$. The total number of boxes of a Young diagram is therefore equal to $n$, and each row contains at most as many boxes as the row above it. A Young tableau is a Young diagram filled with all the integers $1, \dots, n$, one in each boxes. The shape of a Young tableau $Q$, denoted by $\mathrm{shape}(Q)$, is the shape of the associated Young Diagram, it is thus a partition of $n$. There are clearly $n!$ Young tableaux of a given shape $\lambda \vdash n$. A Young tableau is said to be *standard* if the numbers increase along the rows and down the columns.

*Example* 120. In the following figure, the first tableau is standard whereas the second is not.

| | | |
|---|---|---|
| 1 | 2 | 3 |
| 4 | 5 | |
| 6 | | |

| | | |
|---|---|---|
| 1 | 3 | 5 |
| 4 | 2 | |
| 6 | | |

Notice that a standard Young tableau always has 1 in its top-left box, and that the box containing $n$ is necessarily at the end of a row and a column. We denote by $\mathrm{SYT}_n$ the set of all standard Young tableaux of size $n$ and by $\mathrm{SYT}_n(\lambda) = \{Q \in \mathrm{SYT}_n \mid \mathrm{shape}(Q) = \lambda\}$ the set of standard Young tableaux of shape $\lambda$, for $\lambda \vdash n$. By construction, $\mathrm{SYT}_n = \bigsqcup_{\lambda \vdash n} \mathrm{SYT}_n(\lambda)$. Now, a classic result in the representation theory of the symmetric group states that $d_\lambda = |\mathrm{SYT}_n(\lambda)|$ for each $\lambda \vdash n$. The decomposition of Proposition (34) is then refined into:

$$L(\mathfrak{S}_n) \cong \bigoplus_{\lambda \vdash n} \bigoplus_{Q \in \mathrm{SYT}_n(\lambda)} S^{\mathrm{shape}(Q)} \cong \bigoplus_{Q \in \mathrm{SYT}_n} S^{\mathrm{shape}(Q)}. \tag{6.1}$$

Figure 6.1 represents all the standard Young tableaux of size $n = 4$, gathered by shape.

### 6.1.2 The MRA representation and Fourier analysis provide "orthogonal" decompositions of rank information

As explained before, the spaces $S^\lambda$ localize parts of absolute rank information whereas the spaces $H_B$ localize parts of relative rank information. There exists however a connection between the two types of rank information that we detail here. For $k \in \{0, \dots, n\} \setminus \{1\}$, recall that the space $H^k = \bigoplus_{B \subset [\![n]\!],\ |B|=k} H_B$ from Definition 60 is invariant under translations, by Equation (4.11). It is thus also the case of the feature space $\mathbb{H}_n$ and both can be decomposed as a sum of irreducible representations $S^\lambda$:

$$H^k \cong \bigoplus_{\lambda \vdash n} \kappa_\lambda^k S^\lambda \qquad \text{and} \qquad \mathbb{H}_n \cong \bigoplus_{\substack{k=0 \\ k \neq 1}}^{n} \bigoplus_{\lambda \vdash n} \kappa_\lambda^k S^\lambda, \tag{6.2}$$

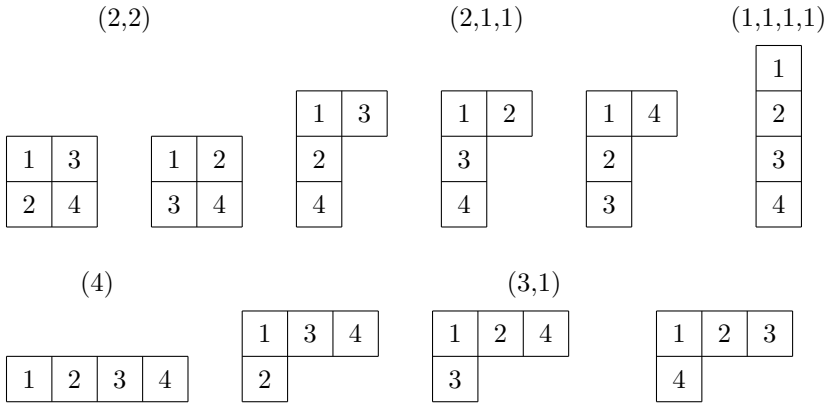



Figure 6.1: Standard Young tableaux of size $n = 4$

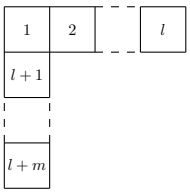
where the $\kappa_\lambda^k$'s are nonnegative integers. Eq. (6.2) means that the space $H^k$ also localizes some absolute rank information, quantified through the multiplicities $\kappa_\lambda^k$ of the $S^\lambda$'s. The connection with the Fourier decomposition of $L(\mathfrak{S}_n)$ is provided in the following proposition.

**Proposition 121** (Representation isomorphism)**.** *The spaces $L(\mathfrak{S}_n)$ and $\mathbb{H}_n$ are isomorphic as representations of $\mathfrak{S}_n$: $L(\mathfrak{S}_n) \cong \mathbb{H}_n$. In particular one has*

$$\sum_{\substack{k=0 \\ k\neq 1}}^{n} \kappa_\lambda^k = d_\lambda \qquad \textit{for all} \qquad \lambda \vdash n.$$

*Proof.* Theorem 47 shows that $\phi_{[\![n]\!]}$ is a linear isomorphism between $\mathbb{H}_n$ and $L(\mathfrak{S}_n)$, and Proposition 61 shows that for any $\tau \in \mathfrak{S}_n$, $T_\tau \phi_{[\![n]\!]} = \phi_{\tau([\![n]\!])} T_\tau = \phi_{[\![n]\!]} T_\tau$. □

The multiplicity $\kappa_\lambda^k$ of each irreducible in $H^k$ can actually be calculated through a combinatorial formula. This is one of the major results established in Reiner et al. (2014). Its statement requires an additional definition. Notice that any standard Young tableau $Q$ contains a unique maximal subtableau of the form




with $1 \leq l \leq n$ and $0 \leq m \leq n - l$. The authors of Reiner et al. (2014) thus define (in the proof of Proposition 6.23) the following quantity:

$$\mathrm{eig}(Q) = \begin{cases} l & \text{if } m \text{ is even,} \\ l-1 & \text{if } m \text{ is odd.} \end{cases} \tag{6.3}$$

This definition enables to specify the Fourier decomposition of each space $H^k$.

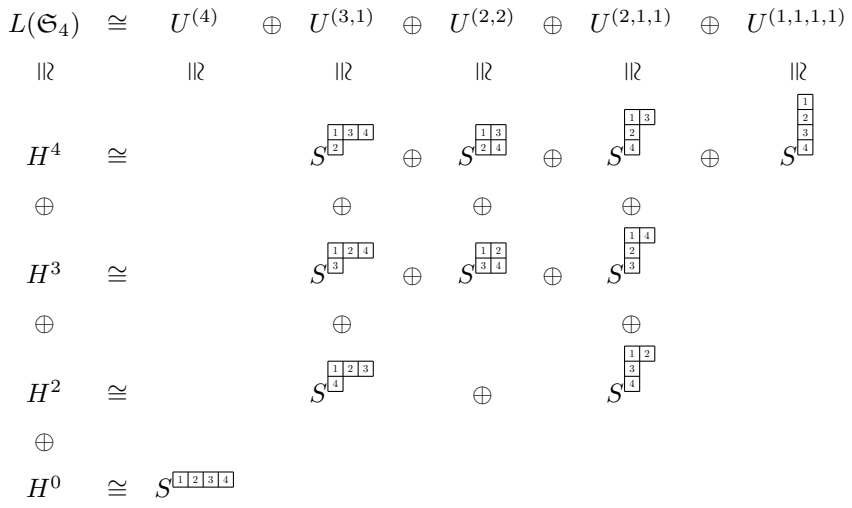


Figure 6.2: Harmonic analysis and MRA decompositions of $L(\mathfrak{S}_4)$

**Theorem 122** (Fourier decomposition of the spaces $H^k$)**.** *For $k \in \{0, \dots, n\} \setminus \{1\}$ and $\lambda \vdash n$, the multiplicity of $S^\lambda$ in $H^k$ is given by $\kappa_\lambda^k = |\{Q \in SYT_n \mid \mathrm{eig}(Q) = n - k\}|$. In other words, the following decomposition holds*

$$H^k \cong \bigoplus_{\substack{Q \in SYT_n \\ \mathrm{eig}(Q) = n-k}} S^{\mathrm{shape}(Q)}.$$

In the notations of Reiner et al. (2014), $H_B = \ker \pi_B$, so that Theorem 122 is a reformulation of their theorem 6.26. It provides a new decomposition of rank information. For $\lambda \vdash n$ we denote by $U^\lambda$ the component $d_\lambda S^\lambda$ in the decomposition of Proposition (34) of $L(\mathfrak{S}_n)$ (it is usually called an isotypic component). Then gathering Proposition (34) with Theorem 122 gives

$$L(\mathfrak{S}_n) \cong \bigoplus_{Q \in \mathrm{SYT}_n} S^{\mathrm{shape}(Q)} \cong \bigoplus_{\lambda \vdash n} U^\lambda \cong \bigoplus_{\substack{k=0 \\ k \neq 1}}^{n} H^k. \tag{6.4}$$

The first decomposition in Equation (6.4) is the full decomposition of $L(\mathfrak{S}_n)$ into irreducible representations, each localizing an "elementary" part of absolute rank information. The second decomposition, into components $U^\lambda$, corresponds to the Fourier decomposition where for each $\lambda \vdash n$, $U^\lambda$ localizes the part of absolute rank information specific to marginals of shape $\lambda$. The last decomposition, into spaces $H^k$, corresponds to the MRA decomposition where for each $k \in \{0, \dots, n\} \setminus \{1\}$, $H^k$ localizes the part of absolute information specific to scale $k$. These different decompositions are illustrated for $n = 4$ in Figure 6.2.

Using the combinatorial formula of Theorem 122 to calculate the multiplicities $\kappa_\lambda^k$, one can obtain some further properties. They are given in the following proposition.

**Proposition 123** (Properties of the multiplicities $\kappa_\lambda^k$)**.** *Let $k \in \{0, \dots, n\} \setminus \{1\}$. One has the following properties:*

1. *The part of absolute rank information of scale $k$ (in terms of MRA) is included in the part of absolute rank information of order $k$ (in terms of Fourier analysis): for any $\lambda \vdash n$ such that $\lambda_1 < n-k$, $\kappa^k_\lambda = 0$.*

2. *There is exactly one copy of the Specht module $S^{(n-1,1)}$ in each of the decompositions of the spaces $H^k$ for $k \in \{2,\dots,n\}$.*

*Proof.* To show Property 1., notice that for $Q \in \mathrm{SYT}_n(\lambda)$, one necessarily has $\mathrm{eig}(Q) \leq \lambda_{\lambda_1}$ by definition (6.3). Thus if $\lambda_1 < n-k$ then $|\{Q \in \mathrm{SYT}_n \mid \mathrm{eig}(Q) = n-k\}| = 0$ and therefore $\kappa^k_\lambda = 0$ by Theorem 122. Property 2. is given by proposition 6.34 from Reiner et al. (2014). □

Notice that the decompositions illustrated by Figure 6.2 satisfy all properties from Propositions 121 and 123.

## 6.2 Alternative construction

In this section we provide some further insights about the alternative embedding $\phi'_{[\![n]\!]}$ considered in Subsection 4.1.3, especially its connection with $\mathfrak{S}_n$-based harmonic analysis, card shuffling and generalized Kendall's tau distances.

### 6.2.1 Alternative embedding of the MRA decomposition into $L(\mathfrak{S}_n)$

We recall that the alternative embedding operator $\phi'_{[\![n]\!]}$ is defined in Equation (4.3) by

$$\phi'_{[\![n]\!]}: \qquad L(\bar{\Gamma}_n) \to L(\mathfrak{S}_n), \qquad F \mapsto \sum_{\pi \in \bar{\Gamma}_n} \frac{|\pi|!}{n!} F(\pi) \mathbb{1}_{\mathfrak{S}_n(\pi)}.$$

We also recall that $\mathfrak{S}_n(\pi)$ is the set of linear extensions of $\pi \in \bar{\Gamma}_n$, which can be seen as the set of full rankings that induce $\pi$ on $c(\pi)$ or as the set of all the possible configurations obtained by shuffling $\pi$ with any ranking $\pi' \in \Gamma([\![n]\!] \setminus c(\pi))$. The former interpretation is behind the approaches introduced in Yu et al. (2002), Kondor and Barbosa (2010) and Sun et al. (2012) and more specifically the empirical ranking model $\widehat{p}_N$ defined in Equation (3.5) is actually equal to $\widehat{p}_N = \frac{1}{N}\sum_{i=1}^N \phi'_{[\![n]\!]}(\delta_{\Pi^{(i)}})$. Huang et al. (2009a) also follows this interpretation and define probabilistic models on $\mathfrak{S}_n$ as linear combinations of elements of the form $\alpha \mathbb{1}_{\mathfrak{S}_n(ij)} + (1-\alpha)\mathbb{1}_{\mathfrak{S}_n(ji)}$ with $1 \leq i < j \leq n$ and $0 \leq \alpha \leq 1$. We show that the part of information contained in these models can be decomposed into components that localize the same part of information as the spaces $H^k$.

For $k \in \{2,\dots,n\}$, we recall that $\Gamma^k_{[\![n]\!]}$ is the set of all incomplete rankings of size $k$. Set $V^0 = \mathbb{R}\mathbb{1}_{\mathfrak{S}_n}$ the space of constant functions on $\mathfrak{S}_n$ and define for $k \in \{2,\dots,n\}$ the space $V^k = \phi'_{[\![n]\!]}(L(\Gamma^k_{[\![n]\!]})) = \mathrm{span}\{\mathbb{1}_{\mathfrak{S}_n(\pi)} \mid \pi \in \Gamma^k_{[\![n]\!]}\}$. One has the following nested sequence of spaces

$$V^0 \subset V^2 \subset \cdots \subset V^n = L(\mathfrak{S}_n).$$

Indeed, $\mathbb{1}_{\mathfrak{S}_n} = \mathbb{1}_{\mathfrak{S}_n(ab)} + \mathbb{1}_{\mathfrak{S}_n(ba)}$ for any distinct $a, b \in [\![n]\!]$, and for $k \in \{2,\dots,n-1\}$, $\pi = \pi_1 \dots \pi_k$ and $a \notin c(\pi)$, one clearly has $\mathbb{1}_{\mathfrak{S}_n(\pi)} = \mathbb{1}_{\mathfrak{S}_n(a\pi_1\dots\pi_k)} + \mathbb{1}_{\mathfrak{S}_n(\pi_1 a\dots\pi_k)} + \dots \mathbb{1}_{\mathfrak{S}_n(\pi_1\dots\pi_k a)}$. We then define the space $W^2$ as the orthogonal supplementary of $V^0$ in $V^2$ and for $k \in \{3,\dots,n\}$ the space $W^k$ as the orthogonal supplementary of $V^{k-1}$ in $V^k$. One thus has $V^0 \overset{\perp}{\oplus} W^2 = V^2$ and

$$V^{k-1} \overset{\perp}{\oplus} W^k = V^k \text{ for all } k \in \{3,\dots,n\} \qquad \text{so that} \qquad L(\mathfrak{S}_n) = V^0 \oplus \bigoplus_{k=2}^n W^k.$$

One would be highly tempted to say that for $k \in \{2, \dots, n\}$, $W^k$ localizes the part of information specific to scale $k$ and $V^k$ localizes the part of information of scales lower or equal than $k$. Fortunately, the following theorem establishes this statement.

**Theorem 124** (Decomposition associated to the alternative embedding)**.** *One has*

$$V^0 = \phi'_{[\![n]\!]}(H^0) \qquad \text{and} \qquad W^k = \phi'_{[\![n]\!]}(H^k) \text{ for all } k \in \{2, \dots, n\}.$$

*In addition, $\phi'_{[\![n]\!]}$ establishes an isomorphism of representations of $\mathfrak{S}_n$ between $\mathbb{H}_n$ and $L(\mathfrak{S}_n)$, so that*

$$V^0 \cong S^{(n)} \qquad \text{and} \qquad W^k \cong H^k \text{ for all } k \in \{2, \dots, n\}.$$

Refer to the Appendix for the proof of Theorem 124. The latter draws the connection between the MRA decomposition and the models that involve the embedding operator $\phi'_{[\![n]\!]}$. In particular it allows to say that for $\pi \in \bar{\Gamma}_n$, the indicator function $\mathbb{1}_{\mathfrak{S}_n(\pi)}$ contains absolute rank information up to scale $|\pi|$. It also naturally recovers some already known results. For instance applying Theorem 122 to $W^2$ gives Proposition 16 in Huang et al. (2009a), or applying Property 1. of Proposition 123 to $W^k$ can be seen as a corollary of Proposition 7 in Kondor and Barbosa (2010).

### 6.2.2 Connection with card shuffling and generalized Kendall's tau distances

The spaces $W^k$ also have an interesting connection with card shuffling, more specifically with random-to-random shuffles. The analysis of card shuffling was introduced in the seminal contributions Aldous and Diaconis (1986) and Bayer and Diaconis (1992). It sees a configuration of a deck of $n$ cards as a permutation of $[\![n]\!]$. The uncertainty about the configuration is then captured by a probability distribution over $\mathfrak{S}_n$. The principle of the analysis of card shuffling is to study the properties of a Markov chain on $\mathfrak{S}_n$ that represents a particular shuffle. The *random-to-random shuffle*, studied in depth in Uyemura-Reyes (2002), consists in picking a card at random from the deck and replacing it at random in the deck. More generally for $k \in \{1, \dots, n-2\}$, the $k$-random-to-random shuffle consists in picking $k$ cards at random from the deck and replacing them at random positions (and in a random order) in the deck. It happens that the transition matrices of the $k$-random-to-random shuffles can be expressed with incomplete rankings.

**Proposition 125** (Connection with card shuffling)**.** *For $k \in \{2, \dots, n-1\}$, the transition matrix $R_k$ of the $(n-k)$-random-to-random shuffling satisfies for any $f \in L(\mathfrak{S}_n)$:*

$$R_k f = (n-k)! \left(\frac{k!}{n!}\right)^2 \sum_{\pi \in \Gamma^k} \left\langle f, \mathbb{1}_{\mathfrak{S}_n(\pi)} \right\rangle \mathbb{1}_{\mathfrak{S}_n(\pi)}.$$

*Proof.* If one picks $n-k$ cards from a configuration $\sigma \in \mathfrak{S}_n$, the configuration of the remaining deck is $\sigma_{|A}$, where $A$ is the subset of $k$ cards that were not picked. Then replacing the $n-k$ cards at random positions and in a random order in the deck can lead to any configuration $\sigma \in \mathfrak{S}_n(\pi)$. The $n-k$-random-to-random shuffle applied to the Dirac function $\delta_\sigma$ therefore decomposes as the sequence of mappings

$$\delta_\sigma \quad \mapsto \quad \frac{1}{\binom{n}{k}} \sum_{A \subset [\![n]\!],\ |A|=k} \delta_{\sigma_{|A}} \quad \mapsto \quad \frac{1}{\binom{n}{k}} \sum_{A \subset [\![n]\!],\ |A|=k} \frac{k!}{n!} \mathbb{1}_{\mathfrak{S}_n(\sigma_{|A})}.$$

Thus for $f = \sum_{\sigma \in \mathfrak{S}_n} f(\sigma) \delta_\sigma$, one has

$$R_k f = \sum_{\sigma \in \mathfrak{S}_n} f(\sigma) \frac{1}{\binom{n}{k}} \sum_{A \subset [\![n]\!],\ |A|=k} \frac{k!}{n!} \mathbb{1}_{\mathfrak{S}_n(\sigma_{|A})} = \frac{1}{\binom{n}{k}} \frac{k!}{n!} \sum_{\pi \in \Gamma^k} \mathbb{1}_{\mathfrak{S}_n(\pi)} \sum_{\sigma \in \mathfrak{S}_n} f(\sigma) \mathbb{I}\{\pi \subset \sigma\}.$$

This concludes the proof. □

By Proposition 125, it is clear that for $k \in \{2, \dots, n-1\}$ the image space of $R_k$ is included in $V^k$ and that its null space contains all spaces $W^j$ for $k < j \leq n$: $\operatorname{Im} R_k \subset V^k$ and $\ker R_k \supset \bigoplus_{j=k+1}^{n} W^j$. These results can actually be refined using the ones from Reiner et al. (2014). The connection is established via the following proposition.

**Proposition 126** (Connection with matrices from Reiner et al. (2014))**.** *Let $k \in \{2, \dots, n-1\}$. For $\sigma, \sigma' \in \mathfrak{S}_n$, $R_k(\sigma, \sigma')$ is proportional to the number of subwords of size $k$ that $\sigma$ and $\sigma'$ have in common:*

$$R_k(\sigma, \sigma') = (n-k)! \left(\frac{k!}{n!}\right)^2 |\{A \subset [\![n]\!] \text{ with } |A| = k \mid \sigma_{|A} = \sigma'_{|A}\}|.$$

*Proof.* Noticing that $\langle \delta_\sigma, \mathbb{1}_{\mathfrak{S}_n(\pi)} \rangle = \mathbb{1}_{\mathfrak{S}_n(\pi)}(\sigma) = \mathbb{I}\{\pi \subset \sigma\}$ for any $\pi \in \bar{\Gamma}_n$ and $\sigma \in \mathfrak{S}_n$, one obtains

$$R_k(\sigma, \sigma') = R_k \delta_{\sigma'}(\sigma) = (n-k)! \left(\frac{k!}{n!}\right)^2 \sum_{\pi \in \Gamma^k} \mathbb{I}\{\pi \subset \sigma'\} \mathbb{I}\{\pi \subset \pi\},$$

which gives the desired result. □

The number $|\{A \subset [\![n]\!] \text{ with } |A| = k \mid \sigma_{|A} = \sigma'_{|A}\}|$ of subwords of size $k \in \{2, \dots, n-1\}$ that $\sigma \in \mathfrak{S}_n$ and $\sigma' \in \mathfrak{S}_n$ have in common is equal to $\operatorname{noninv}_k(\sigma'^{-1}\sigma)$ where $\operatorname{noninv}_k$ is the statistics on $\mathfrak{S}_n$ defined in Reiner et al. (2014). Proposition 126 thus says that the matrix $R_k$ is proportional to the matrix $\nu_{(k,1^{n-k})}$ considered by the authors of Reiner et al. (2014). Now, one of their major results is that these matrices are symmetric positive semidefinite and pairwise commute. They can thus be simultaneously diagonalized and the following result establishes a connection between their eigenspaces and the $W^k$'s.

**Theorem 127** (Null spaces of the matrices $R_k$)**.** *Each of the spaces $V^0$, $W^2$, …, $W^n$ is stable for all the matrices $R_k$ for $k \in \{2, \dots, n-1\}$. It is thus a direct sum of their eigenspaces. In addition, one has*

$$\ker R_k = \bigoplus_{j=k+1}^{n} W^j \qquad \text{for all } k \in \{2, \dots, n-1\}.$$

*Proof.* It is proven in Uyemura-Reyes (2002) that $\dim \ker R_{n-1} = d_n$, the number of derangements on a set of $n$ elements. Since $W^n \subset \ker R_{n-1}$ and $\dim W^n = d_n$ by Theorem 124, one has $\ker R_{n-1} = W^n$ and $\operatorname{Im} R_{n-1} = V^{n-1}$. Now , in Reiner et al. (2014), the authors define in equation (22) the space $V_{n,j} = \ker R_{n-j-1} \cap \operatorname{Im} R_{n-j}$ for $j \in \{1, \dots, n-2\}$. They show that each space $V_{n,j}$ is stable for all matrices $R_k$. They show in addition that for each $j \in \{1, \dots, n-2\}$, $\dim V_{n,j} = \binom{n}{j} d_{n-j}$. For $j = 1$ one then has

$$V_{n,1} = \ker R_{n-2} \cap V^{n-1} \quad \text{and} \quad \ker R_{n-2} \supset W^{n-1} \oplus W^n \quad \text{so that} \quad V_{n,1} \supset W^{n-1}.$$

Again, by Theorem 124, $\dim W^{n-1} = n d_{n-1} = \dim V_{n,1}$ so that $V_{n,1} = W^{n-1}$ and therefore $\ker R_{n-2} = W^{n-1} \oplus W^n$. By induction, one obtains that for all $j \in \{1, \dots, n-2\}$, $V_{n,j} = W^{n-j}$ and $\ker R_{n-j} = \bigoplus_{i=0}^{j-1} W^{n-i}$. This concludes the proof. □

The goal of a shuffle is to mix cards so that the configuration of the deck after several iterations is closest to a purely random configuration. By definition, the component of a probability distribution over $\mathfrak{S}_n$ that lies in the null space of a shuffle is mixed after one iteration (on

average). Theorem 127 therefore says that the space $W^k$ localizes the part of information that is preserved by the $j$-random-to-random shuffles for $1 \leq j \leq n-k$ but mixed by the $j$-random-to-random shuffles for $n-k+1 \leq j \leq n-2$.

Finally, notice that by Proposition 126, $R_2(\cdot,\cdot)$ is proportional to $\binom{n}{2} - d_{KT}(\cdot,\cdot)$ where $d_{KT}$ is the Kendall's tau distance. It thus has the same null space as the matrix $(d_{KT}(\sigma,\sigma'))_{\sigma,\sigma' \in \mathfrak{S}_n}$. More generally for $k \in \{2,\dots,n-1\}$, Proposition 126 gives

$$R_k(\sigma,\sigma') = (n-k)! \left(\frac{k!}{n!}\right)^2 \left(\binom{n}{k} - d^k(\sigma,\sigma')\right),$$

where $d^k(\sigma,\sigma') := |\{A \subset [\![n]\!] \mid |A| = k \text{ and } \sigma_{|A} \neq \sigma_{|A'}\}$ is the number of $k$-wise disagreements between $\sigma$ and $\sigma'$, and can therefore be seen as an extension of the Kendall's tau distance. Hence the matrices of the distances $d^k$ for $k \in \{2,\dots,n-1\}$ pairwise commute and their null spaces are given by Theorem 127.

To conclude this section, we summarize the interpretations that can be given to the spaces $V^0, W^2, \dots, W^n$ and thus to the different scales of the MRA. For $k \in \{2,\dots,n\}$:

- $W^k$ is the space spanned by the $\mathbb{1}_{\mathfrak{S}_n(\pi)}$'s for $\pi \in \Gamma^k$ that localizes the part of absolute rank information specific to scale $k$.
- $W^k$ localizes the part of information preserved by the $j$-random-to-random shuffles for $1 \leq j \leq n-k$ but mixed by the $j$-random-to-random shuffles for $n-k+1 \leq j \leq n-2$.
- $W^k$ localizes the part of additional information captured by the distance $d^k$ compared to $d^{k-1}$.

## 6.3 Absolute rank information at scale 2 and social choice theory

In this section we analyze in particular the part of absolute rank information at scale 2 or in other words the part of information contained in pairwise marginals. We then develop the connection with social choice theory.

### 6.3.1 Decomposition of absolute rank information at scale 2

By Theorem 122, one has the isomorphism $H^2 \cong S^{(n-1,1)} \oplus S^{(n-2,1,1)}$. We give an explicit construction of subspaces of $H^2$ that correspond to this decomposition. First we recall that $H^2 = \bigoplus_{\{a,b\} \subset [\![n]\!]} H_{\{a,b\}}$ with $\dim H_{\{a,b\}} = 1$ for each pair $\{a,b\} \subset [\![n]\!]$, so that $\dim H^2 = \binom{n}{2}$. The following proposition gives an explicit basis for each $H_{\{a,b\}}$ and thus for $H^2$. Its proof is straightforward and left to the reader.

**Proposition 128** (Canonical basis of $H^2$)**.** *For any $a,b \in [\![n]\!]$ with $a \neq b$, the element $x_{a \succ b} = \delta_{ab} - \delta_{ba}$ generates the space $H_{\{a,b\}}$. By convention, we choose for each pair $\{a,b\} \subset [\![n]\!]$ with $a < b$ the element $x_{a \succ b}$ to be the canonical basis of $H_{\{a,b\}}$. The canonical basis of $H^2$ is then given by the family $(x_{a \succ b})_{1 \leq a < b \leq n}$.*

We use the canonical basis introduced in Proposition 128 to construct the two following

subspaces of $H^2$:

$$H_1^2 = \operatorname{span}\left\{ e_a := \sum_{b\in[\![n]\!],\ b\neq a} x_{a\succ b} \;\middle|\; a \in [\![n]\!] \right\}$$
$$\text{and} \qquad H_2^2 = \operatorname{span}\left\{ f_{a,b} := \sum_{c\in[\![n]\!],\ c\notin\{a,b\}} (x_{a\succ b} + x_{b\succ c} + x_{c\succ a}) \;\middle|\; \{a,b\} \subset [\![n]\!] \right\}. \tag{6.5}$$

The following theorem shows that they provide a decomposition of $H^2$ isomorphic to $S^{(n-1,1)} \oplus S^{(n-2,1,1)}$. Its proof is left in Appendix.

**Theorem 129** (Explicit decomposition of $H^2$)**.** *The spaces $H_1^2$ and $H_2^2$ defined in Equation (6.5) satisfy the following properties:*

$$H^2 = H_1^2 \oplus H_2^2 \qquad \textit{with} \qquad H_1^2 \cong S^{(n-1,1)} \qquad \textit{and} \qquad H_2^2 \cong S^{(n-2,1,1)}.$$

It happens that the spaces $H_1^2$ and $H_2^2$ defined in Equation (6.5) appear in several other mathematical constructions and therefore have different interpretation. First they are connected with social choice theory, this is explained in the remaining of this Section. The second connection we detail is with the HodgeRank framework. Introduced in Jiang et al. (2011b), it models a collection of pairwise comparisons as an oriented flow on the graph with vertices $[\![n]\!]$ where two items are linked if the pair appears at least once in the comparisons. The collection of observed pairwise comparisons is the observation design $\mathcal{A}$ of our present setting. The space of edge flows considered in Jiang et al. (2011b) is then equal to the space $\mathbb{H}(\mathcal{A}) = \bigoplus_{\{a,b\}\in\mathcal{A}} H_{\{a,b\}}$. The HodgeRank framework then decomposes any element of this space as the sum of three components: a "gradient flow" that corresponds to globally consistent rankings, a "curl flow" that corresponds to locally inconsistent rankings, and a "harmonic flow", that corresponds to globally inconsistent but locally consistent rankings. The following proposition establishes the connection with the present work.

**Proposition 130** (Connection with HodgeRank)**.** *In the particular case where $\mathcal{A} = \{\{a,b\} \subset [\![n]\!]\}$, the space of edge flows in the HodgeRank framework is equal to $H^2$, the space of gradient flows to $H_1^2$, the space of curl flows to $H_2^2$ and the space of harmonic flows is null. The Hodge decomposition then boils down to $H^2 = H_1^2 \oplus H_2^2$. There is no particular connection in the general case.*

*Proof of Proposition 130.* Following the notations of Jiang et al. (2011b), we denote by $G$ the complete graph on $[\![n]\!]$ and by $K_G$ its clique complex. The space of "edge flows" on $G$ is defined by $C^1(K_G,\mathbb{R}) := \{(X_{i,j})_{i,j} \in \mathbb{R}^{n\times n} \mid X_{i,j} = -X_{j,i}\}$. Identifying index $(i,j)$ with $ij$, one clearly has $C^1(K_G,\mathbb{R}) = H^2$. The HodgeRank decomposition, established by theorem 2 in Jiang et al. (2011b), is then given by

$$H^2 = \operatorname{Im}(\operatorname{grad}) \overset{\perp}{\oplus} \operatorname{Im}(\operatorname{curl}^*) = \operatorname{Im}(\operatorname{grad}) \overset{\perp}{\oplus} \operatorname{Im}(\operatorname{grad})^\perp,$$

where by definition $\operatorname{Im}(\operatorname{grad}) = \{\sum_{1\leq i<j\leq n}(s_i - s_j)x_{i\succ j} \mid s \in \mathbb{R}^n\}$. Now, Lemma 146 shows that for any $s \in \mathbb{R}^n$, an element of the form $\{\sum_{1\leq i<j\leq n}(s_i - s_j)x_{i\succ j}$ is of the form $\sum_{i\in[\![n]\!]} s_i e_i$ and reciprocally. This means that $\operatorname{Im}(\operatorname{grad}) = H_1^2$, which concludes the proof. □

### 6.3.2 Decomposition of $W^2$ into eigenspaces of $R_2$

In this subsection we introduce some notations and results to explain the connection with social choice theory. By Theorem 124, $W^2 = \phi'_{[\![n]\!]}(H^2)$ and $\phi'_{[\![n]\!]}$ is an isomorphism between $\mathbb{H}_n$ and

$L(\mathfrak{S}_n)$. Combined with Theorem 129, this gives the following decomposition of $W^2$:

$$W^2 = \phi'_{[\![n]\!]}(H_1^2) \oplus \phi'_{[\![n]\!]}(H_2^2).$$

For $a, b \in [\![n]\!]$ with $a \neq b$, let $K_{a,b}$, $B_a$ and $C_{(a,b)}$ be the respective embeddings of $x_{a \succ b}$, $e_a$ and $C_{(a,b)}$ with respect to $\phi'_{[\![n]\!]}$:

$$K_{a,b} = \phi'_{[\![n]\!]}(x_{a \succ b}) \qquad B_a = \phi'_{[\![n]\!]}(e_a) \qquad C_{a,b} = \phi'_{[\![n]\!]}(f_{a,b}).$$

One has by construction $B_a = \sum_{c \neq a} K_{a,c}$ and $C_{a,b} = \sum_{c \notin \{a,b\}} (K_{a,b} + K_{b,c} + K_{c,a})$ and the following lemma gives explicit expressions for these elements as functions of $\mathfrak{S}_n$. Its proof is left in Appendix. We recall that $\operatorname{sign}(u) = u/|u|$ for $u \in \mathbb{R} \setminus \{0\}$.

**Lemma 131.** *Let $a, b \in [\![n]\!]$ with $a \neq b$ and $\sigma \in \mathfrak{S}_n$. The following properties hold.*

*(i)* $K_{a,b} = \mathbb{1}_{\mathfrak{S}_n(ab)} - \mathbb{1}_{\mathfrak{S}_n(ba)}$ *or equivalently* $K_{a,b}(\sigma) = \operatorname{sign}(\sigma(b) - \sigma(a))$

*(ii)* $B_a = \sum_{r=1}^{n} (n + 1 - 2r) \mathbb{1}_{\{\sigma(a) = r\}}$ *or equivalently* $B_a(\sigma) = n + 1 - 2\sigma(a)$

*(iii)* $C_{a,b} = nK_{a,b} + B_b - B_a = \sum_{r=1}^{n-1} (n - 2r) \left( \mathbb{1}_{\{\sigma(b) - \sigma(a) = r\}} - \mathbb{1}_{\{\sigma(b) - \sigma(a) = -r\}} \right)$ *or equivalently* $C_{a,b}(\sigma) = n \operatorname{sign}(\sigma(b) - \sigma(a)) + 2(\sigma(a) - \sigma(b))$.

In the following definition, we simply give specific notations for the embeddings of $H_1^2$ and $H_2^2$ with respect to $\phi'_{[\![n]\!]}$.

**Definition 132** (Spaces $\mathscr{B}_n$ and $\mathscr{C}_n$)**.** We define the spaces

$$\mathscr{B}_n = \phi'_{[\![n]\!]}(H_1^2) = \operatorname{span}(B_a)_{1 \leq a \leq n} \qquad \text{and} \qquad \mathscr{C}_n = \phi'_{[\![n]\!]}(H_2^2) = \operatorname{span}(C_{a,b})_{1 \leq a \neq b \leq n}.$$

As $\phi'_{[\![n]\!]}$ establishes an isomorphism of representations of $\mathfrak{S}_n$ between $\mathbb{H}_n$ and $L(\mathfrak{S}_n)$ by Theorem 124, one has $\mathscr{B}_n \cong S^{(n-1,1)}$, so $\dim \mathscr{B}_n = n - 1$, and $\mathscr{C}_n \cong S^{(n-2,1,1)}$ so $\dim \mathscr{C}_n = (n-1)(n-2)/2$. They are given social choice interpretations in the next subsection. Here, we show that they are eingenspaces of $R_2$.

**Theorem 133** (Eigenstructure of $R_2$)**.** *The following table summarizes the full eigenstructure of $R_2$:*

| *Eigenvalue* | *Eigenspace* | *Dimension* |
|---|---|---|
| $1$ | $V^0$ | $1$ |
| $\frac{n+1}{3\binom{n}{2}}$ | $\mathscr{B}_n$ | $n-1$ |
| $\frac{1}{3\binom{n}{2}}$ | $\mathscr{C}_n$ | $\binom{n-1}{2}$ |
| $0$ | $\bigoplus_{k=3}^{n} W^k$ | $n! - \binom{n}{2} - 1$ |

The proof of Theorem 133 is left in Appendix. We point out however that it was already proven in Renteln (2011). Indeed in the latter, the author fully characterizes the eigenstructure of the distance matrix $D$ of the Cayley graph on $\mathfrak{S}_n$ generated by adjacent transpositions (see Subsection 2.4.4 for the definition). Now, it is well known that the metric of this graph is the Kendall's tau distance: $D(\sigma, \sigma') = d_{KT}(\sigma, \sigma')$ for all $\sigma, \sigma' \in \mathfrak{S}_n$. Since $R_2$ is proportional to $\binom{n}{2} J - D$ where $J$ is the $n! \times n!$ with only ones, the results from Renteln (2011) directly apply here. We chose however to construct the eigenspaces $\mathscr{B}_n$ and $\mathscr{C}_n$ with our own notations in order to better interpret and exploit them. For instance the following lemma, proved using their explicit construction, will be useful thereafter. Its proof is left in Appendix.

**Lemma 134.** *The orthogonal projections of an element $p \in L(\mathfrak{S}_n)$ on $\mathscr{B}_n$ and $\mathscr{C}_n$ are given by*

$$p_{\mathscr{B}_n} = \frac{3}{n(n+1)!} \sum_{a \in [\![n]\!]} \langle p, B_a \rangle \, B_a \qquad \text{and} \qquad p_{\mathscr{C}_n} = \frac{3}{n^2.n!} \sum_{\{a,b\} \subset [\![n]\!]} \langle p, C_{a,b} \rangle \, C_{a,b}.$$

### 6.3.3 Connection with social choice theory

In this subsection we study Kemeny rank aggregation. For clarity's sake, we denote by $d$ the Kendall's tau distance on $\mathfrak{S}_n$. We recall that a permutation $\sigma^*$ is a (generalized) Kemeny consensus for a function $p \in L(\mathfrak{S}_n)$ if $\sigma^* = \operatorname{argmin}_{\sigma \in \mathfrak{S}_n} \sum_{\sigma' \in \mathfrak{S}_n} d(\sigma, \sigma') p(\sigma')$ and that the set of such consensuses is denoted by $\mathcal{C}_d(p)$ (see Definition 115). By Proposition 126, this can be reformulated in

$$\mathcal{C}_d(p) = \operatorname*{argmax}_{\sigma \in \mathfrak{S}_n} R_2 p(\sigma). \tag{6.6}$$

Formula (6.6) shows that the Kemeny rank aggregation problem for a function $p \in L(\mathfrak{S}_n)$ is entirely characterized by $R_2 p$. Formally, let $p = p_{\ker R_2} + p_{(\ker R_2)^\perp}$ be the decomposition of $p$ on the null space $\ker R_2$ of $R_2$ and its orthogonal supplementary. One then has $R_2 p = R_2 p_{(\ker R_2)^\perp}$. All information from $p$ filtered by $R_2$ thus does not have any impact on the problem. We therefore call $(\ker R_2)^\perp$ the *effective space* of Kemeny rank aggregation (as in Saari, 2000; Daugherty et al., 2009; Crisman, 2014). By Theorem 133, it admits the following orthogonal linear decomposition

$$(\ker R_2)^\perp = V^0 \oplus \mathscr{B}_n \oplus \mathscr{C}_n.$$

Next we show that the two components $\mathscr{B}_n$ and $\mathscr{C}_n$ have a specific meaning from the point of view of social choice theory.

**Borda space $\mathscr{B}_n$.** As in Saari (2000); Crisman (2014), we call $\mathscr{B}_n$ the *Borda space* (or component). Let us introduce some more definitions to justify this name. For a collection of $N \geq 1$ permutations $(\sigma_1, \dots, \sigma_N) \in \mathfrak{S}_n^N$, the Borda Count is the voting rule that consists in affecting to element $a \in [\![n]\!]$ the score $\sum_{t=1}^N \sigma_t(a)$ and then produce a full ranking of the candidates by sorting them in increasing order of these scores (notice that this does not define a unique output as some candidates may receive the same score). The Borda Count is generalized to the case of a function $p \in L(\mathfrak{S}_n)$ by replacing the score of the candidate $a \in [\![n]\!]$ by $\sum_{\sigma \in \mathfrak{S}_n} \sigma(i) p(\sigma)$ or equivalently by $-\langle p, B_a \rangle$ by Lemma 131. The set $BC(p)$ of the outputs of the Borda Count for $p$ is therefore fully characterized by $p_{\mathscr{B}_n}$. The following proposition refines this observation.

**Proposition 135.** *For $p \in L(\mathfrak{S}_n)$, $BC(p) = \operatorname{argmax}_{\sigma \in \mathfrak{S}_n} p_{\mathscr{B}_n}(\sigma) = \mathcal{C}_d(p_{\mathscr{B}_n})$.*

*Proof.* By Lemma 134 one has,

$$\operatorname*{argmax}_{\sigma \in \mathfrak{S}_n} p_{\mathscr{B}_n}(\sigma) = \operatorname*{argmax}_{\sigma \in \mathfrak{S}_n} \frac{3}{n(n+1)!} \sum_{a \in [\![n]\!]} \langle p, B_a \rangle \, B_a(\sigma) = \operatorname*{argmax}_{\sigma \in \mathfrak{S}_n} \sum_{a \in [\![n]\!]} (-\langle p, B_a \rangle) \sigma(a).$$

Now, it is well known that for any $x_1, \dots, x_n \in \mathbb{R}$, the maximum of $\sum_{i=1}^n x_i \sigma(i)$ is obtained for permutations $\sigma \in \mathfrak{S}_n$ such that $x_{\sigma^{-1}(n)} \geq \cdots \geq x_{\sigma^{-1}(1)}$. This provides the first equality. For the second one, Theorem 133 implies that

$$R_2 p_{\mathscr{B}_n} = \frac{n+1}{3\binom{n}{2}} p_{\mathscr{B}_n} \qquad \text{so that} \qquad \mathcal{C}_d(p_{\mathscr{B}_n}) = \operatorname*{argmax}_{\sigma \in \mathfrak{S}_n} R_2 p_{\mathscr{B}_n}(\sigma) = \operatorname*{argmax}_{\sigma \in \mathfrak{S}_n} p_{\mathscr{B}_n}(\sigma).$$

□

Proposition 135 says that the outputs of the Borda Count for $p$ are exactly the modes of $p_{\mathscr{B}_n}$ and also exactly the Kemeny consensuses of $p_{\mathscr{B}_n}$. This surely justifies the name of "Borda space" for $\mathscr{B}_n$.

**Condorcet space $\mathscr{C}_n$.** The space $\mathscr{C}_n$ is responsible for the *Condorcet paradox* and we thus call it the *Condorcet space*, as in Chandra and Roy (2013); Crisman (2014). The Condorcet paradox says that in an election, it can happen that candidate $a$ wins on average in pairwise duels against candidate $b$, candidate $b$ wins on average in pairwise duels against candidate $c$ but candidate $c$ wins on average in pairwise duels against candidate $a$. Formally for a function $p \in L(\mathfrak{S}_n)$, it can happen that

$$P_{\{a,b\}}(ab) > P_{\{a,b\}}(ba) \qquad P_{\{b,c\}}(bc) > P_{\{b,c\}}(cb) \qquad \text{but} \qquad P_{\{a,c\}}(ca) > P_{\{a,c\}}(ac),$$

where $P_B$ is a short notation for $M_B p$. This situation is all the more paradoxical than $|P_{\{a,b\}}(ab) - P_{\{a,b\}}(ba) + P_{\{b,c\}}(bc) - P_{\{b,c\}}(cb) + P_{\{a,c\}}(ca) - P_{\{a,c\}}(ac)|$ is big. For distinct elements $a, b, c \in [\![n]\!]$ we define

$$\mathrm{Cyc}_{\{a,b,c\}}(p) = |P_{\{a,b\}}(ab) - P_{\{a,b\}}(ba) + P_{\{b,c\}}(bc) - P_{\{b,c\}}(cb) + P_{\{a,c\}}(ca) - P_{\{a,c\}}(ac)|. \quad (6.7)$$

It is easy to see that this quantity does not depend on the ordering of $a, b, c$ and is thus well defined. It measures the "amount of inconsistency" or equivalently the amount of "cyclic votes" in $p$ on $\{a, b, c\}$. The following proposition shows that the total amount of inconsistencies in $p$ is controlled by the orthogonal projection of $p$ on $\mathscr{C}_n$.

**Proposition 136.** *For $p \in L(\mathfrak{S}_n)$,*

$$\frac{n^2 n!}{3(n-2)} \|p_{\mathscr{C}_n}\|_\infty \leq \sum_{\{a,b,c\} \subset [\![n]\!]} \mathrm{Cyc}_{\{a,b,c\}}(p) \leq \frac{n(n-1)(n-2)}{2} \sqrt{3n!} \|p_{\mathscr{C}_n}\|_2.$$

*In particular, $p_{\mathscr{C}_n} = 0$ if and only if $\mathrm{Cyc}_{\{a,b,c\}}(p) = 0$ for all distinct elements $a, b, c \in [\![n]\!]$.*

Refer to the Appendix for the proof of Proposition 136. The latter implies that the component $p_{\mathscr{C}_n}$ is entirely responsible for the presence of cyclic votes in $p$ and therefore for a possible Condorcet paradox with $p$. This certainly justifies the name "Condorcet space" for $\mathscr{C}_n$.

**Borda Count approximation of Kemeny's Rule and pairwise voting inconsistencies** In summary for a function $p \in L(\mathfrak{S}_n)$, Proposition 136 implies that $p_{\mathscr{C}_n}$ is responsible for a possible Condorcet paradox and Proposition 135 says that the Kemeny consensus(es) of $p - p_{\mathscr{C}_n}$ are given by the Borda Count on $p$. These facts were already known (through different results) and used as a justification for removing the Condorcet component from the data or using the Borda Count (Saari and Merlin, 2000; Chandra and Roy, 2013). Here we use the precedent results to obtain a quantitative bound on the error of the Borda Count when seen as an approximation of Kemeny's rule. For $p \in L(\mathfrak{S}_n)$ we denote by $\mathcal{R}_p(\sigma) = \sum_{\pi \in \mathfrak{S}_n} d(\sigma, \pi) p(\pi)$ the approximation cost of a permutation $\sigma \in \mathfrak{S}_n$ and by $\mathcal{R}_p^* = \min_{\sigma \in \mathfrak{S}_n} \mathcal{R}_p(\sigma)$ the optimal cost or equivalently the cost of a Kemeny consensus.

**Theorem 137.** *For $p \in L(\mathfrak{S}_n)$ and $\sigma^{BC} \in BC(p)$,*

$$\mathcal{R}_p\left(\sigma^{BC}\right) - \mathcal{R}_p^* \leq \frac{n-2}{n^2} \sum_{\{a,b,c\} \subset [\![n]\!]} \mathrm{Cyc}_{\{a,b,c\}}(p).$$

*Proof.* By proposition 135, $\sigma^{BC} \in \mathcal{C}_d(p_{\mathscr{B}_n})$. Notice also that $\mathcal{C}_d(q+C) = \mathcal{C}_d(q)$ for any $q \in L(\mathfrak{S}_n)$ and $C \in \mathbb{R}$. We therefore consider that $\sigma^{BC} \in \mathcal{C}_d(p_{\mathscr{B}_n} + p_{V^0})$ where $p_{V^0}$ is the orthogonal projection of $p$ on $V^0$ equal to $(\langle p, \mathbb{1}_{\mathfrak{S}_n} \rangle / n!)\mathbb{1}_{\mathfrak{S}_n}$. Proposition 117 then gives

$$\mathcal{R}_{d,p}\left(\sigma^{BC}\right) - \mathcal{R}_{d,p}^* \leq 2\|D(p - p_{\mathscr{B}_n} - p_{V^0})\|_\infty = 2\|Dp_{\mathscr{C}_n}\|_\infty$$

where $D$ is the $n! \times n!$ matrix defined by $D(\sigma, \sigma') = d(\sigma, \sigma')$ for $\sigma, \sigma' \in \mathfrak{S}_n$. As $D = \binom{n}{2}J - (n!\binom{n}{2}/2)R_2$, where $J$ is the $n! \times n!$ matrix of ones, and $Jp_{\mathscr{C}_n} = 0$, one has $\|Dp_{\mathscr{C}_n}\|_\infty = (n!\binom{n}{2}/2)\|R_2 p_{\mathscr{C}_n}\|_\infty$. The proof is then concluded by injecting Theorem 133 and Proposition 136. □

Theorem 137 provides a theoretical bound on the Kemeny approximation cost of the Borda Count. As one could expect with the previous developments, this bound is proportional to the total amount of inconsistencies $\sum_{\{a,b,c\} \subset [\![n]\!]} \mathrm{Cyc}_{\{a,b,c\}}(p)$ in the function $p$, where $\mathrm{Cyc}_{\{a,b,c\}}(p)$ is defined in Equation (6.7) for distinct elements $a, b, c \in [\![n]\!]$.

This bound has to be compared with the state-of-the-art guarantee established in Coppersmith et al. (2006). The latter contribution shows that the Borda Count is a 5-approximation of Kemeny's rule, that is to say for any $p \in L(\mathfrak{S}_n)$ and $\sigma^{BC} \in BC(p)$,

$$\mathcal{R}_p\left(\sigma^{BC}\right) \leq 5\mathcal{R}_p^*. \tag{6.8}$$

We point out two differences between the bounds. The first is that the bound from Theorem 137 is additive whereas the one from (6.8) is multiplicative as are classic bounds for approximation algorithms. This may mean that the bound from Theorem 137 is not tight enough when $\mathcal{R}_p^*$ is small. Finding a multiplicative equivalent for it would be an interesting future direction. The other difference is that the bound from Theorem 137 *depends* on $p$. This is a major advantage because it means that it can be very small for functions $p$ with few inconsistencies. In particular it is equal to 0 for functions $p \perp \mathscr{C}_n$. A drawback however is that in practice, it requires to be computed on each dataset, and its computation has complexity $O(n^3)$.

Ultimately, the central question remains: which bound is the tightest? As we do not have theoretical results we provide here the results of some numerical experiments, performed both on simulated (probability distributions $p$ drawn uniformly at random on the simplex) and empirical data (probability distributions $p$ from the Sushi dataset). Table 6.1 shows that the classic bound of $4\mathcal{R}_p^*$ for $\mathcal{R}_p\left(\sigma^{BC}\right) - \mathcal{R}_p^*$ is much bigger than the one from Theorem 137. This is due to the fact that in many cases the output of the Borda Count is equal to a Kemeny consensus or has a cost close to the minimal cost, as one can see on the values of the difference $\mathcal{R}_p\left(\sigma^{BC}\right) - \mathcal{R}_p^*$ in Table 6.1. The classic bound then remains big while the bound from Theorem 137 adapts to the data and becomes much smaller.

Another interesting point of view is to assess what would be the constant $C > 0$ such that the bound from Theorem 137 is equal to $C\mathcal{R}_p^*$. Figure 6.3 provides boxplots for the value of the ratio $\frac{n-2}{n^2} \sum_{\{a,b,c\} \subset [\![n]\!]} \mathrm{Cyc}_{\{a,b,c\}}(p)/\mathcal{R}_p^*$ for $p$ drawn uniformly at random on the simplex or from the Sushi dataset. The fact that the ratio seems to be decreasing on random $p$'s and increasing for $p$'s from the Sushi dataset is surprising and may require further investigation. Be that as it may, the ratio remains much smaller than 4 in both cases.

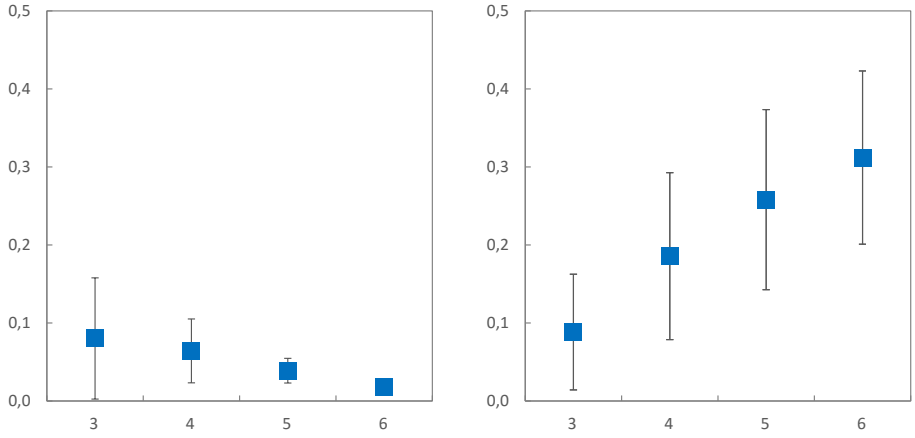
(a) Random $p$

| | $n=3$ | $n=4$ | $n=5$ | $n=6$ |
|---|---|---|---|---|
| $4\mathcal{R}_p^*$ | $4.24 \pm 0.85$ | $10.06 \pm 0.77$ | $18.54 \pm 0.51$ | $29.16 \pm 0.22$ |
| Bound from Theorem 137 | $0.06 \pm 0.04$ | $0.16 \pm 0.09$ | $0.17 \pm 0.07$ | $0.12 \pm 0.04$ |
| $\mathcal{R}_p\left(\sigma^{BC}\right) - \mathcal{R}_p^*$ | $0.02 \pm 0.05$ | $0.03 \pm 0.04$ | $0.03 \pm 0.03$ | $0.02 \pm 0.02$ |

(b) Sushi dataset

| | $n=3$ | $n=4$ | $n=5$ | $n=6$ |
|---|---|---|---|---|
| $4\mathcal{R}_p^*$ | $4.10 \pm 0.94$ | $8.21 \pm 1.46$ | $13.68 \pm 1.99$ | $20.52 \pm 2.42$ |
| Bound from Theorem 137 | $0.08 \pm 0.05$ | $0.35 \pm 0.15$ | $0.83 \pm 0.28$ | $1.54 \pm 0.41$ |
| $\mathcal{R}_p\left(\sigma^{BC}\right) - \mathcal{R}_p^*$ | $0.002 \pm 0.009$ | $0.004 \pm 0.014$ | $0.006 \pm 0.014$ | $0.007 \pm 0.012$ |

Table 6.1: Comparison of Kemeny approximation bounds for the Borda Count




(a) Random $p$ (b) Sushi dataset

Figure 6.3: Ratio $\frac{n-2}{n^2} \sum_{\{a,b,c\} \subset [\![n]\!]} \mathrm{Cyc}_{\{a,b,c\}}(p) / \mathcal{R}_p^*$

# Part III

# Future directions and conclusion

# Chapter 7

# Future directions and conclusion

In this last chapter we describe interesting future directions related this thesis. Section 7.1 proposes some ideas to define regularization procedures for the MRA framework and Section 7.2 discusses extensions of the MRA representation to more general types of ranking data. At last, we give a general conclusion to this thesis in Section 7.3.

**Contents**



## 7.1 Regularization procedures

Here we describe some regularization procedures that one may consider but the list is of course non exhaustive. Our suggestions are based on intuition and analogy with classic regularization procedures on other types of data. Hence they do not come with any theoretical guarantees. Finding a good regularity assumption and the associated regularization procedure in the feature space $\mathbb{H}_n$ largely remains an open problem.

### 7.1.1 Kernel-based smoothing

The most usual way to define a notion of regularity is to say that a function $f$ is regular if "$f(x) \simeq f(y)$" for "$x \simeq y$". In this case, the knowledge of $f(x)$ can be used to infer some knowledge about $f(y)$. Indeed if one has an estimation of $f$ at some point $x$ and assumes that $f$ is regular, he can obtain estimations for points $y \simeq x$. A typical approach is then to regularize an initial estimator by applying a smoothing kernel $K_h$ that will "diffuse" the knowledge of $f(x)$

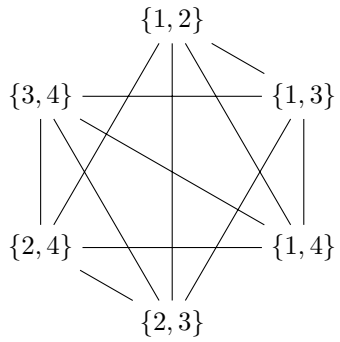



Figure 7.1: Graph on pairs of items for $n = 5$

to points $y$ close to $x$. The parameter $h$ is usually a window parameter that controls both the "speed and the range of the diffusion". As we detailed in Subsection 3.1.5, kernel smoothing for incomplete rankings is already used in Kondor and Barbosa (2010) and Sun et al. (2012). The difference here is that we propose to define kernels on the feature space $\mathbb{H}_n$ instead of the space $L(\mathfrak{S}_n)$.

Here we propose an approach to transpose these ideas for the feature space $\mathbb{H}_n$. By analogy, one wants to say that an element $\mathbf{X} = (X_B)_{B \in \bar{\mathcal{P}}([\![n]\!])}$ is regular if "$X_B \simeq X_{B'}$" for "$B \simeq B'$". The first step is therefore to define relevant meanings for "$X_B \simeq X_{B'}$" and "$B \simeq B'$". We assert that the MRA representation already exploits the consistency assumption to transfer information between included subsets and therefore between different scales. Transferring information between elements $X_B$ and $X_{B'}$ indexed by two subsets of different size is then not relevant. Hence we define a notion of regularity for each subspace $H^k$ and from now on we fix $k \in \{0, 2, \dots, n\}$. First we propose to consider the distance $D_k$ defined for $B, B' \in \bar{\mathcal{P}}([\![n]\!])$ with $|B| = |B'| = k$ by

$$D_k(B, B') = \frac{1}{2}\left(k - |B \cap B'|\right)$$

(the proof that $D_k$ is a distance on $\{B \subset [\![n]\!] \mid |B| = k\}$ is left to the reader). Two subsets $B, B'$ with $|B| = |B'| = k$ thus have distance 1 if they have $k-1$ items in common, 2 if they have $k-2$ items in common, ..., and $k$ if they have no item in common. The distance $D_k$ is also the distance on the graph with set of nodes $\{B \subset [\![n]\!] \mid |B| = k\}$ and where $B$ and $B'$ are connected if they have $k-1$ items in common. An illustration of this graph for $n = 5$ and $k = 2$ is provided on Figure 7.1.

We now define a relevant meaning for "$X_B \simeq X_{B'}$". The difficulty is that for $B \neq B'$, the elements $X_B$ and $X_{B'}$ lie in different spaces and how they should be compared is not obvious. To tackle this problem we propose to send one to the space of the other and then to compare them. For $B, B' \in \bar{\mathcal{P}}([\![n]\!])$ we define the set

$$\mathrm{Bij}(B, B') = \{\tau : B \to B' \text{ bijection } \mid \tau(b) = b \text{ for all } b \in B \cap B'\}.$$

For $\tau \in \mathrm{Bij}(B, B')$ we denote by $\tau(\pi_1 \dots \pi_k) := \tau(\pi_1) \dots \tau(\pi_k)$ and define for $X_B \in H_B$ the element $\tau \cdot X_B := \sum_{\pi \in \Gamma(B)} X_B(\pi)\delta_{\tau(\pi)}$. With a proof similar to the one of 61, it is easy to show that $\tau \cdot X_B \in H_{B'}$. We then say that "$X_B \simeq X_{B'}$" if

$$X_{B'} \simeq \frac{1}{|\,\mathrm{Bij}(B, B')|} \sum_{\tau \in \mathrm{Bij}(B, B')} \tau \cdot X_B \qquad \text{in } H_{B'}.$$

The kernels associated to the regularity assumption "$X_B \simeq X_{B'}$" for "$B \simeq B'$" are then functions $K_h : H^k \to H^k$ defined by

$$K_h : X_B \mapsto \sum_{|B'|=k} \frac{q_h(D_k(B,B'))}{|\operatorname{Bij}(B,B')|} \sum_{\tau \in \operatorname{Bij}(B,B')} \tau \cdot X_B,$$

where $q_h : \mathbb{N} \to \mathbb{R}$ is a nonnegative function such that $\sum_{\pi\in\Gamma^k} K_h X_B(\pi) = \sum_{\pi\in\Gamma(B)} X_B(\pi)$. Since for any $B' \subset [\![n]\!]$ with $|B'| = k$ and $\tau \in \operatorname{Bij}(B,B')$, $\sum_{\pi\in\Gamma^k} \tau \cdot X_B(\pi) = \sum_{\pi\in\Gamma(B)} X_B(\pi)$, the condition on $q_h$ boils down to

$$\sum_{|B'|=k} q_h(D_k(B,B')) = 1 \qquad i.e. \qquad \sum_{j=0}^{k} q_h(j) \binom{k}{j} \binom{n-k}{j} = 1.$$

One can take for instance $q_h(j) = [(h+1)\binom{k}{j}\binom{n-k}{j}]^{-1}$ if $0 \le j \le h$ and 0 otherwise.

### 7.1.2 Penalty minimization and sparsity

Another classic approach to define regularization procedure is through the minimization of a penalty function. One chooses a dissimilarity measure $\Delta$ on $\mathbb{H}_n$, and then defines a regularized version of an initial element $\mathbf{X} \in \mathbb{H}_n$ as the solution of a minimization problem of the form

$$\min_{\mathbf{X}'\in\mathbb{H}_n} \Delta(\mathbf{X},\mathbf{X}') + \lambda\Omega(\mathbf{X}'), \tag{7.1}$$

where $\Omega : \mathbb{H}_n \to \mathbb{R}$ is a penalty function and $\lambda > 0$ is a regularization parameter. As $\mathbb{H}_n$ is constructed as $\bigoplus_{B\in\bar{\mathcal{P}}([\![n]\!])} H_B$, it is natural to define a dissimilarity measure $\Delta$ of the form

$$\Delta(\mathbf{X},\mathbf{X}') = \sum_{B\in\bar{\mathcal{P}}([\![n]\!])} \Delta_B(X_B, X'_B),$$

where for each $B \in \bar{\mathcal{P}}([\![n]\!])$, $\Delta_B$ is a dissimilarity measure on $H_B$. If one takes $\Delta_B = \|\cdot\|_B^2$ then $\Delta := \|\cdot\|^2_{\bar{\Gamma}_n}$, the Euclidean norm on $L(\bar{\Gamma}_n)$. The challenge in this approach lies more in the definition of a "good" penalty function $\Omega$. If one wants to enforce the regularity assumption described previously, one can use the Tikhonov regularization approach and take $\Omega(\mathbf{X}') = \|K_h\mathbf{X}' - \mathbf{X}'\|^2_{\bar{\Gamma}_n}$. The use of a penalty function can also force the solution of (7.1) to be sparse in a certain basis or dictionary. The first challenge is then to define a dictionary where "regular" elements of $\mathbb{H}_n$ should be sparse in. As explained previously, such a dictionary should not contain elements that lie in one single space $H_B$ only. In other words, "regular" elements of $\mathbb{H}_n$ should not have the form $\sum_{i=1}^{s} \alpha_i X_{B_i}$ with a small $s$, where for $i \in \{1,\dots,s\}$, $B_i \in \bar{\mathcal{P}}([\![n]\!])$, $X_{B_i} \in H_{B_i}$ and $\alpha_i \in \mathbb{R}$. Instead, we advocate to define atoms of the form $\mathbf{X}^k_{\mathcal{B}} = \sum_{B\in\mathcal{B}} X_B$ with $\mathcal{B} \subset \{B \subset [\![n]\!] \mid |B| = k\}$ and $X_B \in H_B$ for each $B \in \mathcal{B}$. As an example, we consider for distinct items $a, b \in [\![n]\!]$ the following element (defined in Proposition 128):

$$x_{a\succ b} = \delta_{ab} - \delta_{ba} \in H_{\{a,b\}}.$$

Then one can consider a dictionary with atoms

$$\mathbf{X}^2_{a,B} = \sum_{b\in B} x_{a\succ b} \in \bigoplus_{b\in B} H_{\{a,b\}} \qquad \text{for } a \in [\![n]\!] \text{ and } B \subset [\![n]\!] \setminus \{a\}.$$

Such an atom localizes the part of rank information that says that item $a$ is preferred to each of the items of $B$ in pairwise comparisons in the sense that for $i, j \in [\![n]\!]$ with $i \neq j$,

$$\phi_{\{i,j\}}\mathbf{X}^2_{a,B} = \begin{cases} \delta_{ab} - \delta_{ba} & \text{if } \{i,j\} = \{a,b\} \text{ with } b \in B \\ 0 & \text{otherwise.} \end{cases}$$

### 7.1.3 Fourier band-limited approximation

Another classic regularization procedure is to compute the Fourier transform of a function, truncate it to the low frequencies, and output its inverse. The performance of this procedure for functions on Euclidean spaces stems from the fact that the Fourier spectrum of irregularities is usually localized in high frequencies. Keeping only the low frequencies of the Fourier spectrum of a function $f$ therefore leads to a regularized version of $f$. The analogue of this approach can been applied for functions on the symmetric group, using $\mathfrak{S}_n$-based harmonic analysis (see Huang et al., 2009a; Irurozki et al., 2011). The additional challenge is that "frequencies" are then partitions of $n$ (see Subsection 3.2.5) and thus are not naturally ordered. Fortunately the dominance order (see Definition 32) is a partial order on partitions of $n$ that orders Fourier coefficients by a certain level of "smoothness". Hence the band-limited approximation procedure has been proven to be efficient on real datasets (see Huang et al., 2009a; Irurozki et al., 2011).

This regularization procedure can also be applied to the statistical analysis of incomplete rankings:

1. Compute the wavelet empirical estimator $\widehat{\mathbf{X}} \in \mathbb{H}_n$
2. Apply the procedure to $\phi_{[\![n]\!]}\widehat{\mathbf{X}} \in L(\mathfrak{S}_n)$
3. Compute its wavelet transform to obtain a regularized wavelet estimator $\tilde{\mathbf{X}} \in \mathbb{H}_n$

This procedure is theoretical because it would not lead to tractable computations. For that, one needs to find a way to obtain $\tilde{\mathbf{X}}$ from $\widehat{\mathbf{X}}$ without passing by $\phi_{[\![n]\!]}\widehat{\mathbf{X}}$. We point out this direction however because we assert that this regularization procedure gains a new interpretation when applied to the statistical analysis of incomplete rankings: it allows to regularize small pieces of relative rank information into global parts of absolute rank information. Assume for instance that one observes pairwise comparisons and keeps only absolute rank information of level 1. Besides the piece of rank information of level 0, there are $n(n-1)/2$ potential degrees of freedom in the data, one for the piece of relative rank information related to each pair in $[\![n]\!]$. By contrast, there are only $n-1$ degrees of freedom in the part of absolute rank information localized in the copy of $S^{(n-1,1)}$ that appears in the decomposition of $H^2$ (see Subsection 6.1.2). Keeping only this component therefore allows to enforce the regularity constraints of absolute rank information on the pieces of relative rank information captured by $\widehat{\mathbf{X}}$.

### 7.1.4 Local regularization

In some applications, one is only interested in using an estimator to make local predictions on small subsets of items. One then does not have to regularize the full wavelet empirical estimator $\widehat{\mathbf{X}}$ but can regularize only the coefficients involved in each prediction. For $A \in \mathcal{P}([\![n]\!])$, we recall that the estimation of the marginal $P_A$ of the true ranking model $p$ provided by $\widehat{\mathbf{X}}$ is equal to $\phi_A \sum_{B \in \bar{\mathcal{P}}(A)} \widehat{X}_B$. One therefore only needs to regularize the coefficients $(\widehat{X}_B)_{B \in \bar{\mathcal{P}}(A)} \in \mathbb{H}(\bar{\mathcal{P}}(A))$ to improve the estimation of $P_A$. Thanks to the multiscale nature of $\mathcal{P}([\![n]\!])$, the three aforementioned families of regularization procedures naturally apply to $\mathbb{H}(\bar{\mathcal{P}}(A))$. Notice however that if one wants to apply the the Fourier band-limited approximation procedure, she will have to use the Fourier transform based on $\mathfrak{S}_A$, the group of permutations of $A$. The regularization then will involve "absolute rank information on $A$" and not absolute rank information on $[\![n]\!]$.

The drawbacks of a local regularization procedure is of course that it does not allow to transfer information from subsets of items not included in $A$ to subsets of items included in $A$. The major advantage however is the much lower computational cost: the parameter $n$ that would appear in any of the procedures when regularizing globally becomes $|A|$ when regularizing locally, which is much smaller in practical applications.

## 7.2 Extensions and constructions

In this section we describe some interesting constructions or extensions that could be made on the MRA representation.

### 7.2.1 Exponential models

The same as one can construct exponential parametric models based on Fourier analysis (see Diaconis, 1988), this could be based on the MRA representation. The general approach would be the following. Let $p$ be a probability distribution over $\mathfrak{S}_n$ that we want to model. We assume that $p$ is always positive: $p(\sigma) > 0$ for all $\sigma \in \mathfrak{S}_n$. Instead of decomposing $p$ in the MRA representation, we can decompose $\log p$:

$$\log p = \sum_{B \in \bar{\mathcal{P}}([\![n]\!])} \phi_{[\![n]\!]} \Psi_B \log p.$$

Then to obtain a model with few parameters, one can truncate the decomposition of $\log p$. For instance truncating at scale 2 leads to the following model, by Lemma 71:

$$p_x(\sigma) = \frac{1}{Z_n(x)} e^{\sum_{i=1}^n x_{\sigma_i, \sigma_{i+1}}} \qquad \text{for all} \qquad \sigma \in \mathfrak{S}_n,$$

where $x = (x_{i,j})_{1 \leq i \neq j \leq n} \in \mathbb{R}^{n(n-1)}$ with $x_{j,i} = -x_{i,j}$ for all $i \neq j$ and $Z_n(x)$ is the normalizing factor such that $\sum_{\sigma \in \mathfrak{S}_n} p_x(\sigma) = 1$. The model $p_x$ has $n(n-1)/2$ free parameters. Up to our knowledge, it has never been considered in the literature. Studying it can be an interesting future direction.

### 7.2.2 Extension to the analysis of incomplete rankings with ties

In practical applications, one may observe incomplete rankings with ties. For instance if a user chooses some items $a_1, \ldots, a_k$ among a selection of proposed items $\{a_1, \ldots, a_k, b_1, \ldots, b_l\}$ then one can model her preference by the ranking $a_1, \ldots, a_k \succ b_1, \ldots, b_l$. More generally, incomplete rankings with ties are partial orders of the form

$$a_{1,1}, \ldots, a_{n_1,1} \succ \cdots \succ a_{1,r}, \ldots, a_{n_r,r} \qquad \text{with } r \geq 1 \text{ and } \sum_{i=1}^{r} n_i < n. \tag{7.2}$$

Observations then cannot be represented as incomplete rankings anymore, but as incomplete rankings with ties, and the MRA framework needs to be extended before it can be applied. To do so,observe that an incomplete ranking with ties of the form (7.2) can be seen as a partial ranking on the subset of items $\{a_{1,1}, \ldots, a_{n_r,r}\}$. We therefore propose to extend the MRA framework as follows:

1. Construct an estimator $\widehat{Q}_A$ on each observed subset of items $A$ using any method to analyze partial rankings from the literature
2. Compute the wavelet transforms of all the $\widehat{Q}_A$'s and average them to obtain a wavelet estimator $\tilde{\mathbf{X}}$
3. Perform the task related to the considered application in the feature space $\mathbb{H}_n$ using $\tilde{\mathbf{X}}$ as empirical distribution

Of course, this extended framework needs to be developed for each statistical application with respect to the considered method to analyze partial rankings.

### 7.2.3 Application to label ranking

In label ranking (see Subsection 2.3) from incomplete rankings, one considers a dataset of $N$ IID observations

$$\mathcal{D}_N = ((\mathbf{x}_1, \mathbf{A}_1, \Pi^{(1)}), \dots, (\mathbf{x}_N, \mathbf{A}_N, \Pi^{(N)}))$$

of a random triple $(\mathbf{x}, \mathbf{A}, \Pi)$ where $\mathbf{x}$ is a random variable on an input space $\mathcal{X}$ and $(\mathbf{A}, \Pi)|\mathbf{x} = x$ is drawn from process (3.3) with ranking model $p_x$ and observation design $\nu_x$. The input space is equipped with a structure (typically $\mathcal{X} = \mathbb{R}^d$) and the general principle is to exploit some kind of regularity of the function $x \mapsto p_x$ with respect to this structure. In simple terms, one should have "$p_x \simeq p_{x'}$" for "$x \simeq x'$". We propose to apply the MRA framework to label ranking in a $k$-nearest neighbors regression approach. For a given metric on $\mathcal{X}$, $k \geq 1$ and $x \in \mathcal{X}$ we denote by $\mathcal{D}_N^k(x) = ((\mathbf{A}_{i_1}, \Pi^{(i_1)}), \dots, (\mathbf{A}_{i_k}, \Pi^{(i_k)}))$ the sub-dataset of $\mathcal{D}_N$ of couples $(\mathbf{A}_i, \Pi^{(i)})$ that corresponds to the $k$ closest $\mathbf{x}_i$ to $x$. We then define the $k-nn$ MRA classifier as the wavelet empirical estimator (5.1) for the dataset $\mathcal{D}_N^k(x)$:

$$\widehat{X}_B^k(x) = \frac{1}{\sum_{j=1}^k \mathbb{I}\{B \subset \mathbf{A}_{i_j}\}} \sum_{j=1}^k \Psi_B \delta_{\Pi^{(i_j)}}.$$

### 7.2.4 Extension to an infinite set of elements with features

We now turn to the extension of the MRA framework to rankings on an infinite set of elements with features, say $\mathbb{R}^d$. We thus consider a dataset of $N$ IID observations

$$\mathcal{D}_N = ((\mathbf{A}_1, \Pi^{(1)}), \dots, (\mathbf{A}_N, \Pi^{(N)})),$$

of a random couple $(\mathbf{A}, \Pi)$ except that now $\mathbf{A} = \{\mathbf{x}_1, \dots, \mathbf{x}_k\} \subset \mathbb{R}^d$ is a subset of $\mathbb{R}^d$ with $k = |\mathbf{A}| \geq 2$. In the most general setting, one can still assume that $(\mathbf{A}, \Pi)$ is drawn from (3.3) except that the law $\nu$ of $\mathbf{A}$ is now a probability distribution over finite subsets of $\mathbb{R}^d$ (for instance a spatial Poisson point process) and that $\Pi|\mathbf{A} = A$ is drawn from a probability distribution $P_A$ over $\Gamma(A)$ with $A$ a finite subset of $\mathbb{R}^d$. The infinite family $(P_A)_{A \subset \mathbb{R}^d, |A| < \infty}$ can still be considered as a "ranking model". Consistency assumption $(*)$ does not however have a natural analogue in this setting. In any case, probabilistic modeling in this setting should not lead to an infinite number of parameters. The general assumption is that for two "similar" finite subsets $A, A' \subset \mathbb{R}^d$, the random rankings $\Pi|\mathbf{A} = A$ and $\Pi|\mathbf{A} = A'$ should have "similar laws".

A major difficulty is in the definition of similarity of two finite subsets: not only the subsets should have similar geometric properties but also one is interested in matching the elements of one subset to the ones of the other. For instance one could characterize the geometry of a pair $A = \{x, y\}$ by the line $\mathbb{R}(y - x)$, compare it to another pair $A' = \{x', y'\}$ through the quantity $|\langle y - x, y' - x' \rangle| \|y - x\| \|y' - x'\|$ and if it is big enough assume the correspondence $(x, y) \equiv (x', y')$ if $\langle y - x, y' - x' \rangle > 0$ and $(x, y) \equiv (y', x')$ if $\langle y - x, y' - x' \rangle < 0$. This modeling approach becomes however harder for subsets with more than two elements.

We propose an approach in two steps: first perform a clustering of all the $\mathbf{x}_i$'s that appear in the dataset, then apply the MRA framework on the set of $n$ clusters. More specifically, let $\mathbf{x}_{i,1}, \dots, \mathbf{x}_{i,k_i}$ be the elements of $\mathbf{A}_i$ with for $i = 1, \dots, N$. We propose to first perform a clustering on the dataset $\mathbf{x}_{1,1}, \dots, \mathbf{x}_{i,k_1}, \dots, \mathbf{x}_{N,1}, \dots, \mathbf{x}_{N,k_N}$ into clusters $\mathcal{C}_1, \dots, \mathcal{C}_n$. Each cluster should represent elements that are not comparable or considered as equivalent. The clustering should therefore naturally be based on the structure of $\mathbb{R}^d$ but also exploit the additional part of information provided by the dataset $\mathcal{D}_N$: elements $\mathbf{x}_{i,1}, \dots, \mathbf{x}_{i,k_i}$ in a same observed subset $\mathbf{A}_i$ should all belong to different clusters. These requirements can for instance be satisfied by a

*constrained clustering* algorithm (see Wagstaff et al., 2001). The MRA framework would then naturally apply to the set $[\![n]\!]$ where element $i \in [\![n]\!]$ represents cluster $\mathcal{C}_i$.

## 7.3 Conclusion

In this thesis we have introduced a new representation for ranking data, more specifically a multiresolution analysis (MRA) representation for functions of incomplete rankings. We have established its construction (using recent results from algebraic topology), explained in details the localization properties it offers (comparing with classic MRA and with Fourier analysis on the symmetric group), described an associated Fast Wavelet Transform, developed a framework for statistical applications and shown connections with several other mathematical constructions.

In our point of view, the MRA representation brings many new insights to ranking data analysis and provides a novel, flexible and general framework for many applications. Though we have not demonstrated its power on empirical large-scale settings, we are convinced that highly efficient methods can be constructed with the adapted extensions. This is why we have also provided a global survey of ranking data analysis and detailed many future directions, to best settle this work in the related literature and facilitate its developments.

From a general perspective, I find ranking data analysis fascinating. It provides a unique combination of constraints and possibilities: so many traditional approaches from vector data analysis do not apply to ranking data but at the same time the extraordinarily rich structure of rankings and permutations offers a formidable playground to apply results from various mathematical areas. I hope that the present work will be helpful for the future developments of this field.

To conclude, I believe that the ubiquity of ranking data may not be a mere coincidence. In some sense, classification and clustering, central tasks in machine learning and pattern recognition, are related to probabilistic modeling on equivalence relations. Ranking data analysis on the other hand is related to probabilistic modeling on order relations. Now, these two are the main binary relations considered in mathematics, no other binary relation has attracted similar attention. This is why I believe that the study of ranking data will continue to play a major role in the mathematical sciences.

# Appendices

# Appendix A

# Proofs of Chapter 3

## A.1 Proofs of Section 3.1

*Proof of Lemma 17.* We use the notations introduced in Section 5.1. On the one hand, the number of parameters to store $\mathcal{D}_N$ is obviously bounded by $N$. On the other hand, a dataset $\mathcal{D}_N$ is characterized by the probability distribution $\widehat{\nu}$ and the collection of naive empirical estimators $(\widehat{P_A})_{A\in\widehat{\mathcal{A}}_N}$. The number of parameters required to store $\widehat{\nu}$ is $|\widehat{\mathcal{A}}_N|$, thus at most equal to $|\mathcal{A}|$. The number of parameters required to store $(\widehat{P_A})_{A\in\widehat{\mathcal{A}}_N}$ is equal to $\sum_{A\in\widehat{\mathcal{A}}_N}|\operatorname{supp}(\widehat{P_A})|$, thus at most equal to $\sum_{A\in\mathcal{A}}(|A|!-1)$. Summing these two quantities gives the desired result. □

## A.2 Proofs of Section 3.2

*Proof of Proposition 33.* Property 1 is a direct consequence of Definition 9 of the marginal operators $M_A$. To prove Property 2 we use Theorem 39. The latter stipulates that there exist orthogonal matrices $C_\lambda$ and $C_\mu$ of respective sizes $|\operatorname{Part}_\lambda([\![n]\!])|$ and $|\operatorname{Part}_\mu([\![n]\!])|$ such that

$$M_\lambda f = C_\lambda \left[\bigoplus_{\xi\trianglerighteq\lambda}\bigoplus_{l=1}^{K_{\xi,\lambda}}\widehat{f}(\xi)\right]C_\lambda^\top \qquad \text{and} \qquad M_\mu f = C_\mu \left[\bigoplus_{\xi\trianglerighteq\mu}\bigoplus_{l=1}^{K_{\xi,\mu}}\widehat{f}(\xi)\right]C_\mu^\top.$$

Now, one has $K_{\xi,\lambda}\geq K_{\xi,\mu}$ for any $\xi\vdash n$ because $\mu\trianglerighteq\lambda$ (see for instance Sagan, 2013). We can thus define the linear operator $\Xi_{\lambda,\mu}$ that extracts $\bigoplus_{\xi\trianglerighteq\mu}\bigoplus_{l=1}^{K_{\xi,\mu}}\widehat{f}(\xi)$ from $\bigoplus_{\xi\trianglerighteq\lambda}\bigoplus_{l=1}^{K_{\xi,\lambda}}\widehat{f}(\xi)$. Defining $\mathcal{M}_{\lambda,\mu}$ by

$$\mathcal{M}_{\lambda,\mu}:\mathbb{R}^{|\operatorname{Part}_\lambda([\![n]\!])|\times|\operatorname{Part}_\mu([\![n]\!])|}, \qquad M\mapsto C_\mu\Xi_{\lambda,\mu}\left(C_\lambda^\top M C_\lambda\right)C_\mu^\top$$

then concludes the proof. □

# Appendix B

# Proofs of Chapter 4

## B.1 Proofs of Section 4.1

Lemma 49 is a cornerstone in the construction of the MRA representation. Its proof relies on the exploitation of the combinatorial structure of the embedding operators. Let $\Gamma_n^* := \bar{\Gamma}_n \cup [\![n]\!]$ be the set of all injective words on $[\![n]\!]$, including the words of length 1 of the form $a$, with $a \in [\![n]\!]$. We extend Definition 88 of the concatenation product to words $\pi, \pi' \in \Gamma_n^*$:

$$\pi\pi' := \begin{cases} \pi_1 \dots \pi_{|\pi|}\pi'_1 \dots \pi'_{|\pi'|} & \text{if } c(\pi) \cap c(\pi') = \emptyset, \\ 0 & \text{if } c(\pi) \cap c(\pi') \neq \emptyset. \end{cases}$$

The following lemma gives a combinatorial expression for the embedding operator.

**Lemma 138.** *Let $\pi \in \Gamma_n$ and $A \in \mathcal{P}([\![n]\!])$ such that $c(\pi) \subset A$. Then one has*

$$\phi_A \delta_\pi = \frac{1}{(|A| - |\pi| + 1)!} \sum_{\substack{A_1, A_2 \subset A \\ A_1 \sqcup A_2 = A \setminus c(\pi)}} \sum_{\substack{\omega \in \Gamma(A_1) \\ \omega' \in \Gamma(A_2)}} \delta_{\omega\pi\omega'}.$$

*Proof.* The proof only consists in noticing that

$$\{\sigma \in \Gamma(A) \mid \pi \sqsubset \sigma\} = \{\omega\pi\omega' \mid (\omega, \omega') \in \Gamma(A_1) \times \Gamma(A_2) \text{ with } A_1 \sqcup A_2 = A \setminus c(\pi)\}.$$

□

Lemma 49 then relies on the two following lemmas. The proof of the first one is straightforward and left to the reader.

**Lemma 139.** *Let $A, A \subset [\![n]\!]$ be two disjoint subsets and $(\pi, \pi') \in \Gamma(A) \times \Gamma(A')$. Then for any $B \in \mathcal{P}([\![n]\!])$ such that $B \cap A \neq \emptyset$ and $B \cap A' \neq \emptyset$ one has*

$$(\pi\pi')_{|B} = \pi_{|B\cap A}\pi'_{|B\cap A'}.$$

**Lemma 140.** *For any $n, r, s \in \mathbb{N}$, one has the identity*

$$\sum_{k=0}^{n} \binom{k+r}{r} \binom{n-k+s}{s} = \binom{n+r+s+1}{n}$$

*Proof.* Denote the sum by $S_n(r,s)$. By Pascal's rule, one has

$$\begin{aligned} S_n(r+1,s) &= \sum_{k=0}^{n} \binom{k+r+1}{k}\binom{n-k+s}{s} \\ &= \binom{n+s}{s} + \sum_{k=1}^{n} \binom{k+r}{k}\binom{n-k+s}{s} + \sum_{k=1}^{n} \binom{k+r}{k-1}\binom{n-k+s}{s} \\ &= \sum_{k=0}^{n} \binom{k+r}{k}\binom{n-k+s}{s} + \sum_{k=0}^{n-1} \binom{k+r+1}{k}\binom{n-1-k+s}{s} \\ &= S_n(r,s) + S_{n-1}(r+1,s). \end{aligned}$$

One thus has $S_n(r+1,s) - S_{n-1}(r+1,s) = S_n(r,s)$ and, noticing that $S_0(r,s) = 1$ for all $r,s \in \mathbb{N}$, one obtains by a telescoping sum

$$S_n(r+1,s) = \sum_{k=0}^{n} S_k(r,s).$$

The identity is now proven by induction on $r$ using the well-known identity

$$\sum_{j=k}^{n} \binom{j}{k} = \binom{n+1}{k+1} \tag{B.1}$$

(it can be proven by induction on $n$ with Pascal's rule). For $r = 0$ one has

$$S_n(0,s) = \sum_{k=0}^{n} \binom{n-k+s}{s} = \sum_{j=s}^{n+s} \binom{j}{s} = \binom{n+s+1}{s+1},$$

which satisfies the identity. Assuming the identity true for all $k \leq r$, one has

$$S_n(r+1,s) = \sum_{k=0}^{n} \binom{k+r+s+1}{r+s+1} = \sum_{j=r+s+1}^{n+r+s+1} \binom{j}{r+s+1} = \binom{n+r+s+2}{r+s+2},$$

where the last equality also stems from identity (B.1). This concludes the proof. □

*Proof of Lemma 49.* Let $A, B, C \in \mathcal{P}([\![n]\!])$ such that $A \cup B \subset C$ and $\pi \in \Gamma(A)$. We need to prove that $M_B \phi_C \delta_\pi = \phi_B M_{A\cap B} \delta_\pi$. Lemma 138 gives on the one hand

$$\phi_B M_{A\cap B}\delta_\pi = \phi_B \delta_{\pi_{|A\cap B}} = \frac{1}{(|B| - |A\cap B| + 1)!} \sum_{\substack{B_1,B_2\subset B \\ B_1\sqcup B_2 = B\setminus A}} \sum_{\substack{\omega\in\Gamma(B_1) \\ \omega'\in\Gamma(B_2)}} \delta_{\omega\pi_{|A\cap B}\omega'}$$

and on the other hand

$$(|C| - |A| + 1)! M_B \phi_C \delta_\pi = M_B \sum_{\substack{C_1,C_2\subset C \\ C_1\sqcup C_2 = C\setminus A}} \sum_{\substack{\omega\in\Gamma(C_1) \\ \omega'\in\Gamma(C_2)}} \delta_{\omega\pi\omega'} = \sum_{\substack{C_1,C_2\subset C \\ C_1\sqcup C_2 = C\setminus A}} \sum_{\substack{\omega\in\Gamma(C_1) \\ \omega'\in\Gamma(C_2)}} \delta_{(\omega\pi\omega')_{|B}}.$$

Now, by Lemma 139, one has for any $C_1, C_2 \subset C$ such that $C_1 \sqcup C_2 = C \setminus A$ and $(\omega,\omega') \in \Gamma(C_1)\times\Gamma(C_2)$,

$$(\omega\pi\omega')_{|B} = \omega_{|B\cap C_1}\, \pi_{|A\cap B}\, \omega'_{|B\cap C_2}.$$

Therefore, doing the change of variables $B_1 := C_1 \cap B$, $B_2 := C_2 \cap B$, $\upsilon := \omega_{|B\cap C_1}$ and $\upsilon' := \omega'_{|B\cap C_2}$, one obtains

$$M_B \phi_C \delta_\pi = \frac{1}{(|C|-|A|+1)!} \sum_{\substack{B_1,B_2\subset B \\ B_1\sqcup B_2 = B\setminus A}} \sum_{\substack{\upsilon\in\Gamma(B_1) \\ \upsilon'\in\Gamma(B_2)}} c(B_1,B_2,\upsilon,\upsilon')\delta_{\upsilon\pi_{|A\cap B}\upsilon'},$$

where the coefficient $c(B_1, B_2, \upsilon, \upsilon')$ is given by

$$\begin{aligned}
c(B_1,B_2,\upsilon,\upsilon') &= \sum_{\substack{C_1,C_2\subset C \\ C_1\sqcup C_2 = C\setminus A}} \sum_{\substack{\omega\in\Gamma(C_1) \\ \omega'\in\Gamma(C_2)}} \mathbb{I}\{C_1\cap B = B_1, C_2\cap B = B_2, \omega_{|B_1}=\upsilon, \omega'_{|B_2}=\upsilon'\} \\
&= \sum_{\substack{C_1,C_2\subset C \\ C_1\sqcup C_2 = C\setminus A}} \mathbb{I}\{C_1\cap B = B_1, C_2\cap B = B_2\} \frac{|C_1|!}{|B_1|!}\frac{|C_2|!}{|B_2|!} \\
&= \frac{|C_1|!}{|B_1|!}\frac{|C_2|!}{|B_2|!} \sum_{k=0}^{|C|-|A\cup B|} (k+|B_1|)!(|C|-|A\cup B|-k+|B_2|)! \\
&= (|C|-|A\cup B|)! \sum_{k=0}^{|C|-|A\cup B|} \binom{k+|B_1|}{|B_1|}\binom{|C|-|A\cup B|-k+|B_2|}{|B_2|} \\
&= (|C|-|A\cup B|)! \binom{|C|-|A\cup B|+|B_1|+|B_2|+1}{|C|-|A\cup B|},
\end{aligned}$$

where the last equality is given by Lemma 140 for $n := |C| - |A \cup B|$, $r := |B_1|$ and $s := |B_2|$. The proof is concluded by noticing that for $B_1, B_2 \subset B$ such that $B_1 \sqcup B_2 = B \setminus A$, $|B_1|+|B_2| = |B| - |A \cap B|$ and $|A \cup B| - |B_1| - |B_2| = |A|$, so that

$$\binom{|C|-|A\cup B|+|B_1|+|B_2|+1}{|C|-|A\cup B|} = \frac{(|C|-|A|+1)!}{(|C|-|A\cup B|)!(|B|-|A\cap B|+1)!}.$$

□

## B.2 Proofs of Section 4.2

*Proof of Proposition 61.* We prove Property 3 first then Property 1 then Property 2.

- **Property 3.** Since $|\sigma(B)| = |B|$, $\dim H_{\sigma(B)} = \dim H_B$. It is thus sufficient to prove that $T_\sigma(H_B) \subset H_{\sigma(B)}$. For $F \in L(\Gamma(B))$, it is clear that $T_\sigma F = \sum_{\pi\in\Gamma(B)} F(\pi)\delta_{\sigma(\pi)} \in L(\Gamma(\sigma(B)))$. We just need to show that $M_C T_\sigma F = 0$ for any $C \in \bar{\mathcal{P}}(\sigma(B)) \setminus \{\sigma(B)\}$ or equivalently $M_{\sigma(B')} T_\sigma F = 0$ for any $B' \in \bar{\mathcal{P}}(B) \setminus \{B\}$. This is proven by noticing that for any $\pi \in \Gamma(B)$, $(\sigma(\pi))_{|\sigma(B')} = \sigma(\pi_{|B'})$.
- **Property 1.** For $\pi \in \bar{\Gamma}_n$ and $\pi' \in \Gamma(A)$ it is clear that $\pi \sqsubset \pi' \Rightarrow \tau(\pi) \sqsubset \tau(\pi')$. Hence, the mapping $\pi' \mapsto \tau(\pi')$ being injective, $\tau(\{\pi' \in \Gamma(A) \mid \pi \sqsubset \pi'\}) = \{\pi' \in \Gamma(\tau(A)) \mid \tau(\pi) \sqsubset \pi'\}$ and one has

$$\begin{aligned}
(|A|-|\pi|+1)! T_\tau \phi_A \delta_\pi &= T_\tau \mathbb{1}_{\{\pi'\in\Gamma(A)\mid \pi\sqsubset\pi'\}} \\
&= \mathbb{1}_{\tau(\{\pi'\in\Gamma(A)\mid \pi\sqsubset\pi'\})} \\
&= \mathbb{1}_{\{\pi'\in\Gamma(\tau(A))\mid \tau(\pi)\sqsubset\pi'\}} \\
&= (|A|-|\pi|+1)!\phi_{\tau(A)}\delta_{\tau(\pi)}.
\end{aligned}$$

- **Property 2.** Let $B' \in \bar{\mathcal{P}}([\![n]\!])$ with $B \subset B'$ and $F \in L(\Gamma(B'))$. By Theorem 57 one has $F = \phi_{B'} \sum_{B \in \bar{\mathcal{P}}(B')} \Psi_B F$. Applying the operator $T_\tau$ and using the previous result one obtains
$$T_\tau F = T_\tau \phi_{B'} \sum_{B \in \bar{\mathcal{P}}(B')} \Psi_B F = \phi_{\tau(B')} \sum_{B \in \bar{\mathcal{P}}(B')} T_\tau \Psi_B F$$
where for each $B \in \bar{\mathcal{P}}(B')$, $T_\tau \Psi_B F \in H_{\tau(B)}$ by Property 3. On the other hand, applying Theorem 57 to $T_\tau F \in L(\Gamma(\tau(B')))$ gives
$$T_\tau F = \phi_{\tau(B')} \sum_{B \in \bar{\mathcal{P}}(\tau(B'))} \Psi_B T_\tau F = \phi_{\tau(B')} \sum_{B \in \bar{\mathcal{P}}(B')} \Psi_{\tau(B)} T_\tau F.$$
The uniqueness of the MRA decomposition concludes the proof.

□

*Proof of Lemma 63.* Let $B \in \mathcal{P}(A)$. By Definition 44 of the embedding operator
$$\phi_A X_\emptyset = \frac{X_\emptyset(\bar{0})}{|A|!} \mathbb{1}_{\Gamma(A)} \qquad \text{and} \qquad \phi_A X_B = \sum_{\pi \in \Gamma(B)} \frac{X_B(\pi)}{(|A| - |B| + 1)!} \mathbb{1}_{\{\pi' \in \Gamma(A), \pi \sqsubset \pi'\}}.$$
One thus has $\|\phi_A X_\emptyset\|_{A,r}^r = (|X_\emptyset(\bar{0})|/|A|!)^r \times |A|! = |X_\emptyset(\bar{0})|^r/(|A|!^{r-1})$ and
$$\|\phi_A X_B\|_{A,r}^r = \sum_{\pi \in \Gamma(B)} \left| \frac{X_B(\pi)}{(|A| - |B| + 1)!} \right|^r \times (|A| - |B| + 1)! = \frac{\|X_B\|_{B,r}^r}{(|A| - |B| + 1)!^{r-1}},$$
giving Property 1. Property 2 is a direct consequence, using Theorem 57 to have $M_{A'} \phi_A X_B = \phi_{A'} X_B \mathbb{I}\{B \subset A'\}$ for all $A', B \in \mathcal{P}(A)$. □

## B.3 Proofs of Section 4.3

*Proof of Proposition 82.* First, Lemma 74 implies two simplifications:

- First, for $k \in \{2, \dots, n\}$, the coefficients $(\alpha_B(\pi, \pi'))_{\pi, \pi' \in \Gamma(B)}$ are obtained directly from the $(\alpha_{[\![k]\!]}(\pi, \pi'))_{\pi, \pi' \in \Gamma([\![k]\!])}$ for all $B \subset [\![n]\!]$ with $|B| = k$.
- Second, for $B = \{b_1, \dots, b_k\} \in \mathcal{P}([\![n]\!])$ with $b_1 < \dots < b_k$, the coefficients $(\alpha_B(\pi, \pi'))_{\pi' \in \Gamma(B)}$ are obtained directly from the $(\alpha_B(b_1 \dots b_k, \pi'))_{\pi' \in \Gamma(B)}$ for any $\pi \in \Gamma(B)$.

With the precedent simplifications, one only needs to compute and store the $j!$ coefficients $(\alpha_{[\![j]\!]}(12 \dots j, \pi))_{\pi \in \Gamma([\![j]\!])}$ for each $j \in \{2, \dots, k\}$. These coefficients are computed using the recursive formula from Theorem 72. Let $j \in \{2, \dots, k\}$. If all coefficients the $\alpha_{[\![j']\!]}(12 \dots j', \pi)$ for $\pi \in \Gamma([\![j']\!])$ and $2 \le j' \le j - 1$, it is easy to see that the computation of each $\alpha_{[\![j]\!]}(12 \dots j, \pi)$ for $\pi \in \Gamma([\![j]\!])$ then has complexity bounded by $\binom{j}{2}$. The global complexity of the computation of the coefficients $(\alpha_{\{1,\dots,j\}}(12 \dots j, \pi))_{\pi \in \Gamma([\![j]\!]), 2 \le j \le k}$ is therefore bounded by
$$\sum_{j=2}^{k} \binom{j}{2} j! \le \frac{k-1}{2} \sum_{j=2}^{k} [(j+1)! - j!] \le \frac{1}{2} k^2 k!.$$
This establishes Proposition 82. □

## B.4 Proofs of Section 4.4

*Proof of Theorem 91's Property 1.* First, we define the content of a chain $X \in L(\bar{\Gamma}_n)$ as the union of contents of the rankings in its support: $c(X) = \bigcup_{\pi \in \mathrm{supp}(X)} c(\pi)$. Notice then that for $X, Y \in L(\bar{\Gamma}_n)$ with $c(X) \cap c(Y) = \emptyset$, and for $A \in \mathcal{P}([\![n]\!])$,

$$M_A(XY) = M_A(X)M_AY \qquad \text{and} \qquad M_A(X \diamond Y) = M_A(X) \diamond M_AY. \tag{B.2}$$

In particular $M_A(X \diamond Y) = 0$ if $A \subset c(X)$. Let now $B \in \mathcal{P}([\![n]\!])$ and $\tau \in \mathrm{Der}(B)$. We need to show that for all $B' \subsetneq B$, $M_{B'}X_\tau = 0$. Let $B' \subsetneq B$ and $\tau = \gamma_1 \dots \gamma_r$ be the standard cycle form of $\tau$. By definition of Algorithm 3, $X_\tau = X_{\gamma_1} \dots X_{\gamma_r}$. Applying Equation (B.2) thus gives

$$M_{B'}X_\tau = M_{B' \cap \mathrm{supp}(\gamma_1)}X_{\gamma_1} \dots M_{B' \cap \mathrm{supp}(\gamma_r)}X_{\gamma_r}.$$

Now, by definition of $\mathrm{Der}(B)$, $\{\mathrm{supp}(\gamma_1), \dots, \mathrm{supp}(\gamma_r)\}$ is a partition of $B$. Thus there exists at least an index $i \in \{1, \dots, r\}$ such that $B' \cap \mathrm{supp}(\gamma_i) \subsetneq \mathrm{supp}(\gamma_i)$. Let $b \in \mathrm{supp}(\gamma_i) \setminus B'$. Since $\gamma_i$ is a cycle, its support contains at least two elements, and thus $X_{\gamma_i}$ contains a product $b \diamond Y$ or $Y \diamond b$ with $b \notin c(Y)$. Then $M_{B' \cap \mathrm{supp}(\gamma_i)}X_{\gamma_i}$ contains the product $M_{c(Y) \cap B' \cap \mathrm{supp}(\gamma_i)}(b \diamond Y)$. Now the important fact is that $b \notin c(Y) \cap B' \cap \mathrm{supp}(\gamma_i)$, so that

$$M_{c(Y) \cap B' \cap \mathrm{supp}(\gamma_i)}(b \diamond Y) = 0 \qquad \text{by Equation (B.2).}$$

This implies that $M_{B' \cap \mathrm{supp}(\gamma_i)}X_{\gamma_i} = 0$ and so that $M_{B'}X_\tau = 0$, which concludes the proof. □

# Appendix C

# Proofs of Chapter 5

## C.1 Proofs of Section 5.2

The proof of Theorem 106 requires the following Lemmas.

**Lemma 141** (Conditional variance properties)**.** *Let $d \geq 1$, $X$ a random variable on $\mathbb{R}^d$ with $\mathbb{E}[\|X\|^2] < \infty$ and $\mathcal{B}$ a sigma-algebra included in the Borel $\sigma$-algebra on $\mathbb{R}^d$. We define the conditional variance of $X$ with respect to $\mathcal{B}$ by $\mathbb{V}ar[X|\mathcal{B}] = \mathbb{E}\left[\|X - \mathbb{E}[X|\mathcal{B}]\|^2|\mathcal{B}\right]$.*

1. *For all vector $a \in \mathbb{R}^d$,*

$$\mathbb{E}\left[\|X-a\|^2\right] = \mathbb{E}\left[\|\mathbb{E}[X|\mathcal{B}]-a\|^2\right] + \mathbb{E}\left[\mathbb{V}ar[X|\mathcal{B}]\right].$$

2. *For all matrix $M \in \mathbb{R}^{d'\times d}$ and $i \in \{1,\dots,d'\}$,*

$$\mathbb{V}ar\left[MX_i|\mathcal{B}\right] = \sum_{j=1}^{d} M_{i,j}^2 \mathbb{V}ar[X_j|\mathcal{B}] + \sum_{1\leq j\neq k\leq d} M_{i,j}M_{i,k}\mathbb{C}ov[X_j, X_k|\mathcal{B}_N^\nu],$$

*where $\mathbb{C}ov[Y,Z|\mathcal{B}] = \mathbb{E}[(Y-\mathbb{E}[Y|\mathcal{B}])(Z-\mathbb{E}[Z|\mathcal{B}])|\mathcal{B}]$ for any real random variables $Y, Z$.*

*Proof.* As $\mathbb{E}[\|X\|^2] < \infty$, $\mathbb{E}[X|\mathcal{B}]$ is equal to the orthogonal projection of $X$ onto the space $\mathbb{L}^2(\mathcal{B},\mathbb{R}^2)$ of squared-integrable $\mathcal{B}$-measurable random vectors on $\mathbb{R}^d$. Property 1 is thus the Pythagorean theorem applied to $X-a$. For Property 2, one has for $M \in \mathbb{R}^{d'\times d}$ and $i \in \{1,\dots,d'\}$

$$\mathbb{V}ar\left[MX_i\right] = \mathbb{V}ar\left[\sum_{j=1}^{d} M_{i,j}X_j\right] = \sum_{j=1}^{d} M_{i,j}^2 \mathbb{V}ar\left[X_j\right] + \sum_{1\leq j\neq k\leq d} M_{i,j}M_{i,k}\mathbb{C}ov[X_j, X_k|\mathcal{B}_N^\nu].$$

A simple sum inversion then concludes the proof. □

**Lemma 142.** *For $A \in \mathcal{P}([\![n]\!])$ and $\pi, \pi' \in \Gamma(A)$ with $\pi \neq \pi'$,*

$$\mathbb{E}\left[\widehat{P_A}\middle|\mathcal{B}_N^\nu\right] = \mathbb{I}\{A \in \widehat{\mathcal{A}}_N\}P_A$$

$$\mathbb{V}ar\left[\widehat{P_A}(\pi)\middle|\mathcal{B}_N^\nu\right] = \frac{\mathbb{I}\{A \in \widehat{\mathcal{A}}_N\}}{\widehat{N}_A}P_A(\pi)(1-P_A(\pi))$$

*and*

$$\mathbb{C}ov\left[\widehat{P_A}(\pi), \widehat{P_A}(\pi')\middle|\mathcal{B}_N^\nu\right] = -\frac{\mathbb{I}\{A \in \widehat{\mathcal{A}}_N\}}{\widehat{N}_A}P_A(\pi)P_A(\pi').$$

*Proof.* Let $A \in \mathcal{P}([\![n]\!])$ and $\pi \in \Gamma(A)$. Rewriting the empirical estimator $\widehat{P_A}$ as

$$\widehat{P_A} = \frac{\mathbb{I}\{A \in \widehat{\mathcal{A}}_N\}}{\widehat{N}_A} \sum_{i \in \widehat{I}_A} \delta_{\Pi^{(i)}},$$

it is easy to see that $\widehat{P_A}(\pi)|\mathcal{B}_N^\nu$ has the law of $\mathbb{I}\{A \in \widehat{\mathcal{A}}_N\}\widehat{Z}_\pi$, where $\widehat{N}_A\widehat{Z}_\pi$ is a Binomial variable of parameters $\widehat{N}_A$ and $P_A(\pi)$. This gives the expectation and the variance. For the covariance one has if $A \in \widehat{\mathcal{A}}_N$

$$\widehat{P_A}(\pi)\widehat{P_{A'}}(\pi) = \frac{1}{\widehat{N}_A^2} \sum_{1 \le i,j \le \widehat{N}_A} \mathbb{I}\{\Pi^{(i)} = \pi, \Pi^{(j)} = \pi'\} = \frac{1}{\widehat{N}_A^2} \sum_{1 \le i \ne j \le \widehat{N}_A} \mathbb{I}\{\Pi^{(i)} = \pi\}\mathbb{I}\{\Pi^{(j)} = \pi'\}$$

where conditionally to $\mathcal{B}_N^\nu$, $\mathbb{I}\{\Pi^{(i)} = \pi\}$ and $\mathbb{I}\{\Pi^{(j)} = \pi'\}$ are independent Bernoulli variables of respective parameters $P_A(\pi)$ and $P_A(\pi')$. Thus

$$\begin{aligned}
\mathbb{C}ov\left[\widehat{P_A}(\pi), \widehat{P_A}(\pi') \middle| \mathcal{B}_N^\nu\right] &= \mathbb{E}\left[\widehat{P_A}(\pi)\widehat{P_A}(\pi') \middle| \mathcal{B}_N^\nu\right] - \mathbb{E}\left[\widehat{P_A}(\pi) \middle| \mathcal{B}_N^\nu\right] \mathbb{E}\left[\widehat{P_A}(\pi') \middle| \mathcal{B}_N^\nu\right] \\
&= \mathbb{I}\{A \in \widehat{\mathcal{A}}_N\} \frac{\widehat{N}_A - 1}{\widehat{N}_A} P_A(\pi)P_A(\pi') - \mathbb{I}\{A \in \widehat{\mathcal{A}}_N\} P_A(\pi)P_A(\pi') \\
&= -\frac{\mathbb{I}\{A \in \widehat{\mathcal{A}}_N\}}{\widehat{N}_A} P_A(\pi)P_A(\pi').
\end{aligned}$$

□

**Lemma 143.** *For $B \in \mathcal{P}([\![n]\!])$,*

$$\mathbb{P}\left[B \in \mathcal{P}(\widehat{\mathcal{A}}_N)\right] = 1 - (1 - \nu[\mathcal{Q}(B)])^N \qquad \text{and} \qquad \mathbb{E}\left[\frac{\mathbb{I}\{B \in \mathcal{P}(\widehat{\mathcal{A}}_N)\}}{\sum_{A \in \mathcal{Q}(B)} \widehat{N}_A}\right] \le \frac{2}{\nu[\mathcal{Q}(B)](N+1)},$$

*where $\nu[\mathcal{S}] = \sum_{A \in \mathcal{S}} \nu(A)$ for any collection of subsets $\mathcal{S} \subset \mathcal{P}([\![n]\!])$.*

*Proof.* Let $B \in \mathcal{P}([\![n]\!])$. It is easy to see that the random variable

$$\widehat{Z}_{N,B} = \sum_{A \in \mathcal{Q}(B)} \widehat{N}_A = \sum_{A \in \mathcal{Q}(B)} \sum_{i \in \widehat{I}_A} \mathbb{I}\{\mathbf{A}_i = A\} = \sum_{i=1}^N \mathbb{I}\{\mathbf{A}_i \in \mathcal{Q}(B)\}$$

is binomial with parameters $N$ and $\nu[\mathcal{Q}(B)]$ and that $\mathbb{I}\{B \in \mathcal{P}(\widehat{\mathcal{A}}_N)\} = \mathbb{I}\{\widehat{Z}_{N,B} \ge 1\}$. One thus has

$$\mathbb{P}\left[B \in \mathcal{P}(\widehat{\mathcal{A}}_N)\right] = 1 - \mathbb{P}\left[\widehat{Z}_{N,B} = 0\right] = 1 - (1 - \nu[\mathcal{Q}(B)])^N$$

and

$$\mathbb{E}\left[\frac{\mathbb{I}\{B \in \mathcal{P}(\widehat{\mathcal{A}}_N)\}}{\sum_{A \in \mathcal{Q}(B)} \widehat{N}_A}\right] = \mathbb{E}\left[\frac{\mathbb{I}\{\widehat{Z}_{N,B} \ge 1\}}{\widehat{Z}_{N,B}}\right] \le \mathbb{E}\left[\frac{2}{\widehat{Z}_{N,B} + 1}\right],$$

where the last inequality stems from the fact that $z + 1 \le 2z$ for all $z \ge 1$. Now, Chao and Strawderman (1972) provides the following closed-form expression, for a binomial random variable $Z$ of parameters $(n, p)$,

$$\mathbb{E}\left[\frac{1}{Z+1}\right] = \frac{1 - (1-p)^{n+1}}{p(n+1)}.$$

Injecting it concludes the proof. □

*Proof of Theorem 106.* Property 1 being a direct consequence of Property 2, we only show the latter. First, the Cauchy-Schwarz inequality combined with Lemma 63 and the fact that $\widehat{X}_\emptyset = \bar{0} = \Psi_\emptyset p$ lead to the following bound.

$$\mathcal{E}_N\left(\widehat{Q}^{MRA}\right) \leq \sum_{B\in\mathcal{P}(\mathcal{A})}\left(\sum_{A\in\mathcal{Q}(B)}\frac{2^{|A|}\nu(A)}{(|A|-|B|+1)!}\right)\mathbb{E}\left[\|\widehat{X}_B - \Psi_B p\|_B^2\right]. \tag{C.1}$$

Let now $B\in\mathcal{P}(\mathcal{A})$. Lemma 141 provides the decomposition

$$\mathbb{E}\left[\|\widehat{X}_B - \Psi_B p\|_B^2\right] = \mathbb{E}\left[\|\mathbb{E}[\widehat{X}_B|\mathcal{B}_N^\nu] - \Psi_B p\|_B^2\right] + \mathbb{E}\left[\mathbb{V}ar[\widehat{X}_B|\mathcal{B}_N^\nu]\right], \tag{C.2}$$

so that we can bound independently $\mathbb{E}\left[\|\mathbb{E}[\widehat{X}_B|\mathcal{B}_N^\nu] - \Psi_B p\|_B^2\right]$ and $\mathbb{E}\left[\mathbb{V}ar[\widehat{X}_B|\mathcal{B}_N^\nu]\right]$. For that, we first rewrite the wavelet empirical estimator as

$$\widehat{X}_B = \sum_{A\in\mathcal{Q}(B)}\frac{\widehat{N}_A}{\sum_{A'\in\mathcal{Q}(B)}\widehat{N}_{A'}}\Psi_B\widehat{P_A}.$$

The first part of Lemma 142 then gives

$$\begin{aligned}
\mathbb{E}\left[\widehat{X}_B\middle|\mathcal{B}_N^\nu\right] &= \sum_{A\in\mathcal{Q}(B)}\frac{\widehat{N}_A}{\sum_{A'\in\mathcal{Q}(B)}\widehat{N}_{A'}}\Psi_B\mathbb{E}\left[\widehat{P_A}|\mathcal{B}_N^\nu\right]\\
&= \sum_{A\in\mathcal{Q}(B)}\frac{\widehat{N}_A}{\sum_{A'\in\mathcal{Q}(B)}\widehat{N}_{A'}}\mathbb{I}\{A\in\widehat{\mathcal{A}}_N\}\Psi_B P_A\\
&= \sum_{A\in\mathcal{Q}(B)}\frac{\widehat{N}_A}{\sum_{A'\in\mathcal{Q}(B)}\widehat{N}_{A'}}\Psi_B p\\
&= \mathbb{I}\{B\in\mathcal{P}(\widehat{\mathcal{A}}_N)\}\Psi_B p
\end{aligned}$$

so that

$$\mathbb{E}\left[\|\mathbb{E}[\widehat{X}_B|\mathcal{B}_N^\nu] - \Psi_B p\|_B^2\right] = \mathbb{P}\left[B\notin\mathcal{P}(\widehat{\mathcal{A}}_N)\right]\Psi_B p. \tag{C.3}$$

The second part of Lemma 142 gives first

$$\mathbb{V}ar\left[\widehat{X}_B\middle|\mathcal{B}_N^\nu\right] = \sum_{A\in\mathcal{Q}(B)}\left(\frac{\widehat{N}_A}{\sum_{A'\in\mathcal{Q}(B)}\widehat{N}_{A'}}\right)^2\mathbb{V}ar\left[\Psi_B\widehat{P_A}\middle|\mathcal{B}_N^\nu\right]$$

because the variables $(\widehat{N}_A/\sum_{A'\in\mathcal{Q}(B)}\widehat{N}_{A'})\Psi_B\widehat{P_A}$ for $A\in\mathcal{Q}(B)$ are two-by-two independent conditionally to $\mathcal{B}_N^\nu$. Now, for $A\in\mathcal{Q}(B)$ and $\pi\in\Gamma(B)$, one has by Definition 70 of the alpha coefficients

$$\Psi_B\widehat{P_A}(\pi) = \sum_{\pi'\in\Gamma(B)}\alpha_B(\pi,\pi')M_B P_A(\pi') = \sum_{\pi'\in\Gamma(B)}\alpha_B(\pi,\pi')\sum_{\substack{\sigma\in\Gamma(A)\\ \sigma_{|B}=\pi'}}p(\pi') = \sum_{\sigma\in\Gamma(A)}\alpha_B(\pi,\sigma_{|B})P_A(\sigma).$$

Lemmas 141 and 142 then give

$$\begin{aligned}
\mathbb{V}ar\left[\Psi_B\widehat{P_A}(\pi)\middle|\mathcal{B}_N^\nu\right] &= \sum_{\sigma\in\Gamma(A)}\alpha_B(\pi,\sigma_{|B})^2\mathbb{V}ar\left[\Psi_B\widehat{P_A}(\pi)\middle|\mathcal{B}_N^\nu\right] \\
&\qquad + \sum_{\substack{\sigma,\sigma'\in\Gamma(A)\\ \sigma\neq\sigma'}}\alpha_B(\pi,\sigma_{|B})\alpha_B(\pi,\sigma'_{|B})\mathbb{C}ov\left[\widehat{P_A}(\sigma),\widehat{P_A}(\sigma')\middle|\mathcal{B}_N^\nu\right] \\
&= \frac{\mathbb{I}\{A\in\widehat{\mathcal{A}}_N\}}{\widehat{N}_A}\left(\sum_{\sigma\in\Gamma(A)}\alpha_B(\pi,\sigma_{|B})^2P_A(\sigma)(1-P_A(\sigma))\right. \\
&\qquad \left. - \sum_{\substack{\sigma,\sigma'\in\Gamma(A)\\ \sigma\neq\sigma'}}\alpha_B(\pi,\sigma_{|B})\alpha_B(\pi,\sigma'_{|B})P_A(\sigma)P_A(\sigma')\right) \\
&= \frac{\mathbb{I}\{A\in\widehat{\mathcal{A}}_N\}}{\widehat{N}_A}\left(\Psi_B^2p(\pi)-\Psi_Bp(\pi)^2\right),
\end{aligned}$$

where $\Psi_B^2 : L(\bar{\Gamma}_n)\to L(\Gamma(B))$ is the linear operator defined by $\Psi_B^2F(\pi)=\sum_{\sigma\in\bar{\Gamma}_n}\alpha_B^2(\pi,\sigma_{|B})F(\sigma)$ for any $F\in L(\bar{\Gamma}_n)$. One thus has

$$\mathbb{V}ar\left[\widehat{X}_B\middle|\mathcal{B}_N^\nu\right] = \sum_{A\in\mathcal{Q}(B)}\left(\frac{\widehat{N}_A}{\sum_{A'\in\mathcal{Q}(B)}\widehat{N}_{A'}}\right)^2\frac{\mathbb{I}\{A\in\widehat{\mathcal{A}}_N\}}{\widehat{N}_A}\left(\|\Psi_B^2p\|_{B,1}-\|\Psi_Bp\|_{B,2}^2\right)$$

so that

$$\mathbb{E}\left[\mathbb{V}ar\left[\widehat{X}_B|\mathcal{B}_N^\nu\right]\right] = \mathbb{E}\left[\frac{\mathbb{I}\{B\in\mathcal{P}(\widehat{\mathcal{A}}_N)\}}{\sum_{A\in\mathcal{Q}(B)}\widehat{N}_A}\right]\left(\|\Psi_B^2p\|_{B,1}-\|\Psi_Bp\|_2^{B,2}\right). \tag{C.4}$$

Combining Equations (C.1), (C.2), (C.3), (C.4) together with Lemma 143 concludes the proof. □

# Appendix D

# Proofs of Chapter 6

## D.1 Proofs of Section 6.2

The proof of Theorem 124 relies on the properties of the embedding operator $\phi'_A$, given by the following lemma. For $A \in \mathcal{P}([\![n]\!])$ and $\pi' \in \Gamma^{|A|}$, we define the operator $T_{A\to\pi'} : L(\bar{\Gamma}_n) \to \bar{\Gamma}_n$ that maps the Dirac function of a ranking $\pi \in \bar{\Gamma}_n$ to the Dirac function of the ranking obtained by replacing $\pi_{|A}$ by $\pi'$ if $A \subset c(\pi)$ or to 0 otherwise.

**Lemma 144.** *Let $A \in \mathcal{P}([\![n]\!])$ and $\pi \in \Gamma(A)$. The following properties hold.*

1. *For all $A', C \in \mathcal{P}([\![n]\!])$ such that $A \subset A' \subset C$,*
$$\phi'_C \delta_\pi = \phi'_C \phi'_{A'} \delta_\pi.$$
2. *For all $B, C \in \mathcal{P}([\![n]\!])$ such that $A \cup B \subset C$,*
$$M_B \phi'_C \delta_\pi = M_B \phi'_{A\cup B} \delta_\pi.$$
3. *For all $B \in \mathcal{P}([\![n]\!])$,*
$$M_B \phi'_{A\cup B} \delta_\pi = \sum_{\substack{A_1 \subset A\setminus B \\ B_1 \subset B\setminus A \\ |A_1|=|B_1|}} \lambda_{|B_1|} \sum_{\pi' \in \Gamma(B_1)} \phi'_B M_{(A\cap B)\sqcup B_1} T_{A_1\to\pi'} \delta_\pi,$$
*where $\lambda_t = (|A|!|B|!)/(|A\cup B|!(|A\cap B|+t)!)$ for any $t \in \mathbb{N}$.*
4. *For all $\tau \in \mathfrak{S}_n$*
$$T_\tau \phi'_A = \phi'_{\tau(A)} T_\tau$$

*Proof.* We prove the properties in the order.
1. Let $A', C \in \mathcal{P}([\![n]\!])$ such that $A \subset A' \subset C$. One has

$$\phi'_C \phi'_{A'} \delta_\pi = \frac{|A|!}{|A'|!} \phi'_C \sum_{\substack{\pi' \in \Gamma(A') \\ \pi \subset \sigma}} \delta_{\pi'} = \frac{|A|!}{|A'|!} \frac{|A'|!}{|C|!} \sum_{\substack{\pi' \in \Gamma(A') \\ \pi \subset \pi'}} \sum_{\substack{\sigma \in \Gamma(C) \\ \pi' \subset \sigma}} \delta_\sigma = \frac{|A|!}{|C|!} \sum_{\substack{\sigma \in \Gamma(C) \\ \pi \subset \sigma}} \delta_\sigma = \phi'_C \delta_\pi.$$

2. Let $B, C \in \mathcal{P}([\![n]\!])$ such that $A \cup B \subset C$. By definition of the marginal operator and by Property 1., one has

$$M_B \phi'_C \delta_\pi = M_B M_{A\cup B} \phi'_C \phi'_{A\cup B} \delta_\pi.$$

Now, for any $A' \in \mathcal{P}(C)$ and $\pi' \in \Gamma(A')$, it is clear that $M_{A'} \phi'_C \delta_{\pi'} = \delta_{\pi'}$. Applied to $A \cup B$, this concludes the proof of Property 2.

3. This is certainly the longest part of the proof. We introduce two new operators. First, the deletion operator

$$\varrho_a : \delta_\pi \mapsto \delta_{\pi\setminus\{a\}} \qquad \text{for } a \in c(\pi),$$

where $\pi \setminus \{a\}$ is the ranking obtained by deleting the item $a$ in $\pi$. Second, the insertion operator

$$\varrho_b^* : \delta_\pi \to \sum_{i=1}^{|\pi|+1} \delta_{\pi \lhd_i b} \qquad \text{for } b \notin c(\pi),$$

where $\pi \lhd_i b$ is the ranking obtained by inserting item $b$ at the $i^{th}$ position. Then for $A' \in \mathcal{P}(A)$ with $A \setminus A' = \{a_1, \dots, a_r\}$, and $B$ such that $A \subset B$ with $B \setminus A = \{b_1, \dots, b_s\}$, one has

$$M_{A'} \delta_\pi = \varrho_{a_1} \dots \varrho_{a_r} \delta_\pi \qquad \text{and} \qquad \phi'_B \delta_\pi = \frac{|A|!}{|B|!} \varrho_{b_1}^* \dots \varrho_{b_s}^*.$$

Property 3. is then equivalent for any $B \in \mathcal{P}([\![n]\!])$ to

$$\varrho_{a_1} \dots \varrho_{a_r} \varrho_{b_1}^* \dots \varrho_{b_s}^* \delta_\pi = \sum_{k=0}^{\min(r,s)} \sum_{\substack{A_1 \sqcup \{a_{i_1}, \dots, a_{i_{r-k}}\} = \{a_1, \dots, a_r\} \\ B_1 \sqcup \{b_{j_1}, \dots, b_{j_{s-k}}\} = \{b_1, \dots, b_s\} \\ |A_1| = |B_1| = k}} \sum_{\pi' \in \Gamma(B_1)} \varrho_{b_{j_1}}^* \dots \varrho_{b_{j_{s-k}}}^* \varrho_{a_{i_1}} \dots \varrho_{a_{i_{r-k}}} T_{A_1 \to \pi'} \delta_\pi, \quad \text{(D.1)}$$

where $\{a_1, \dots, a_r\} = A \setminus B$ and $\{b_1, \dots, b_s\} = B \setminus A$. We prove Formula (D.1) in three steps. First for $r = s = 1$, one has

$$\varrho_a \varrho_b^* \delta_\pi = \sum_{i=1}^{|A|+1} \varrho_a \delta_{\pi \lhd_i b} = \delta_{\pi \lhd_1 b \setminus \{a\}} + \dots + \delta_{\pi \lhd_{\pi(a)} b \setminus \{a\}} + \delta_{\pi \lhd_{\pi(a)+1} b \setminus \{a\}} + \dots + \delta_{\pi \lhd_1 b \setminus \{a\}}.$$

The ranking $\pi \lhd_{\pi(a)} b \setminus \{a\}$ is the ranking obtained by inserting $b$ at the left of $a$ in $\pi$ and then by deleting $a$. The ranking $\pi \lhd_{\pi(a)+1} b \setminus \{a\}$ is the ranking obtained by inserting $b$ at the right of $a$ in $\pi$ and then by deleting $a$. It is clear that they are both equal to the ranking $\pi_{\{a\}\to b}$ obtained by changing $a$ to $b$ in $\pi$. Hence one has

$$\varrho_a \varrho_b^* \delta_\pi = \varrho_b^* \varrho_a \delta_\pi + T_{\{a\}\to b} \delta_\pi$$

and Formula (D.1) is satisfied. We now show by induction on $s \in \{1, \dots, |B \setminus A|\}$ that

$$\varrho_a \varrho_{b_1}^* \dots \varrho_{b_s}^* \delta_\pi = \varrho_{b_1}^* \dots \varrho_{b_s}^* \varrho_a \delta_\pi + \sum_{i=1}^{s} \varrho_{b_1}^* \dots \varrho_{b_{i-1}}^* \varrho_{b_{i+1}}^* \dots \varrho_{b_s}^* T_{\{a\}\to b_i} \delta_\pi. \quad \text{(D.2)}$$

Notice that for any $A_1 \subsetneq A \setminus B$, $\pi' \in \Gamma(B_1)$ with $B_1 \subset B \setminus A$, $a \in A \setminus (A_1 \sqcup B)$ and $b \in B \setminus (A \sqcup B_1)$ one clearly has

$$\varrho_a T_{A_1 \to \pi'} \delta_\pi = T_{A_1 \to \pi'} \varrho_a \delta_\pi \qquad \varrho_b^* T_{A_1 \to \pi'} \delta_\pi = T_{A_1 \to \pi'} \varrho_b^* \delta_\pi. \quad \text{(D.3)}$$

Therefore, assuming (D.2) true for $s \leq |B \setminus A| - 1$, one has

$$\begin{aligned}
\varrho_a \varrho^*_{b_1} \dots \varrho^*_{b_{s+1}} \delta_\pi &= \varrho_a \varrho^*_{b_1} \dots \varrho^*_{b_s} \left( \varrho^*_{b_{s+1}} \delta_\pi \right) \\
&= \varrho^*_{b_1} \dots \varrho^*_{b_s} \varrho_a \left( \varrho^*_{b_{s+1}} \delta_\pi \right) + \sum_{i=1}^{s} \varrho^*_{b_1} \dots \varrho^*_{b_{i-1}} \varrho^*_{b_{i+1}} \dots \varrho^*_{b_s} T_{\{a\} \to b_i} \left( \varrho^*_{b_{s+1}} \delta_\pi \right) \\
&= \varrho^*_{b_1} \dots \varrho^*_{b_{s+1}} \varrho_a \delta_\pi + \varrho^*_{b_1} \dots \varrho^*_{b_s} T_{\{a\} \to b_{s+1}} + \sum_{i=1}^{s} \varrho^*_{b_1} \dots \varrho^*_{b_{i-1}} \varrho^*_{b_{i+1}} \dots \varrho^*_{b_{s+1}} T_{\{a\} \to b_i} \delta_\pi \\
&= \varrho^*_{b_1} \dots \varrho^*_{b_{s+1}} \varrho_a \delta_\pi + \sum_{i=1}^{s+1} \varrho^*_{b_1} \dots \varrho^*_{b_{i-1}} \varrho^*_{b_{i+1}} \dots \varrho^*_{b_{s+1}} T_{\{a\} \to b_i} \delta_\pi,
\end{aligned}$$

which concludes the proof of (D.2). At last, we show (D.1) by induction on $r \in \{1, \dots, |A \setminus B|\}$. Assuming it true for $r \leq |A \setminus B| - 1$, one has

$$\begin{aligned}
&\varrho_{a_1} \dots \varrho_{a_{r+1}} \varrho^*_{b_1} \dots \varrho^*_{b_s} \delta_\pi \\
&= \varrho_{a_{r+1}} \left[ \varrho_{a_1} \dots \varrho_{a_r} \varrho^*_{b_1} \dots \varrho^*_{b_s} \delta_\pi \right] \\
&= \varrho_{a_{r+1}} \left[ \sum_{k=0}^{\min(r,s)} \sum_{\substack{A_1 \sqcup \{a_{i_1}, \dots, a_{i_{r-k}}\} = \{a_1, \dots, a_r\} \\ B_1 \sqcup \{b_{j_1}, \dots, b_{j_{s-k}}\} = \{b_1, \dots, b_s\} \\ |A_1| = |B_1| = k}} \sum_{\pi' \in \Gamma(B_1)} \varrho^*_{b_{j_1}} \dots \varrho^*_{b_{j_{s-k}}} \varrho_{a_{i_1}} \dots \varrho_{a_{i_{r-k}}} T_{A_1 \to \pi'} \delta_\pi \right].
\end{aligned}$$

If $r \leq s$, Equations (D.2) and (D.3) give

$$\begin{aligned}
&\varrho_{a_1} \dots \varrho_{a_{r+1}} \varrho^*_{b_1} \dots \varrho^*_{b_s} \delta_\pi \\
&= \sum_{k=0}^{r} \sum_{\substack{A_1 \sqcup \{a_{i_1}, \dots, a_{i_{r-k}}\} = \{a_1, \dots, a_r\} \\ B_1 \sqcup \{b_{j_1}, \dots, b_{j_{s-k}}\} = \{b_1, \dots, b_s\} \\ |A_1| = |B_1| = k}} \sum_{\pi' \in \Gamma(B_1)} \Big[ \varrho^*_{b_{j_1}} \dots \varrho^*_{b_{j_{s-k}}} \varrho_{a_{r+1}} \varrho_{a_{i_1}} \dots \varrho_{a_{i_{r-k}}} T_{A_1 \to \pi'} \delta_\pi \\
&\qquad + \sum_{i=1}^{s-k} \varrho^*_{b_{j_1}} \dots \varrho^*_{b_{j_{i-1}}} \varrho^*_{b_{j_{i+1}}} \dots \varrho^*_{b_{j_{s-k}}} T_{\{a_{r+1}\} \to b_{j_i}} \varrho_{a_{i_1}} \dots \varrho_{a_{i_{r-k}}} T_{A_1 \to \pi'} \delta_\pi \Big] \\
&= \sum_{k=0}^{r} \sum_{\substack{A_1 \sqcup \{a_{i_1}, \dots, a_{i_{r+1-k}}\} = \{a_1, \dots, a_{r+1}\} \\ B_1 \sqcup \{b_{j_1}, \dots, b_{j_{s-k}}\} = \{b_1, \dots, b_s\} \\ |A_1| = |B_1| = k \\ a_{r+1} \notin A_1}} \sum_{\pi' \in \Gamma(B_1)} \varrho^*_{b_{j_1}} \dots \varrho^*_{b_{j_{s-k}}} \varrho_{a_{i_1}} \dots \varrho_{a_{i_{r+1-k}}} T_{A_1 \to \pi'} \delta_\pi \\
&+ \sum_{k=1}^{r+1} \sum_{\substack{A_1 \sqcup \{a_{i_1}, \dots, a_{i_{r-k}}\} = \{a_1, \dots, a_{r+1}\} \\ B_1 \sqcup \{b_{j'_1}, \dots, b_{j'_{s-k-1}}\} = \{b_1, \dots, b_s\} \\ |A_1| = |B_1| = k+1 \\ a_{r+1} \in A_1}} \sum_{\pi' \in \Gamma(B_1)} \varrho^*_{b_{j'_1}} \dots \varrho^*_{b_{j'_{s-k-1}}} \varrho_{a_{i_1}} \dots \varrho_{a_{i_{r-k}}} T_{A_1 \to \pi'} \delta_\pi \\
&= \sum_{k=0}^{r+1} \sum_{\substack{A_1 \sqcup \{a_{i_1}, \dots, a_{i_{r+1-k}}\} = \{a_1, \dots, a_{r+1}\} \\ B_1 \sqcup \{b_{j_1}, \dots, b_{j_{s-k}}\} = \{b_1, \dots, b_s\} \\ |A_1| = |B_1| = k}} \sum_{\pi' \in \Gamma(B_1)} \varrho^*_{b_{j_1}} \dots \varrho^*_{b_{j_{s-k}}} \varrho_{a_{i_1}} \dots \varrho_{a_{i_{r+1-k}}} T_{A_1 \to \pi'} \delta_\pi.
\end{aligned}$$

If $s < r$, Equations (D.2) and (D.3) give

$$\begin{aligned}
&\varrho_{a_1} \cdots \varrho_{a_{r+1}} \varrho^*_{b_1} \cdots \varrho^*_{b_s} \delta_\pi \\
&= \sum_{k=0}^{s-1} \sum_{\substack{A_1 \sqcup \{a_{i_1}, \dots, a_{i_{r-k}}\} = \{a_1, \dots, a_r\} \\ B_1 \sqcup \{b_{j_1}, \dots, b_{j_{s-k}}\} = \{b_1, \dots, b_s\} \\ |A_1| = |B_1| = k}} \sum_{\pi' \in \Gamma(B_1)} \Big[ \varrho^*_{b_{j_1}} \cdots \varrho^*_{b_{j_{s-k}}} \varrho_{a_{r+1}} \varrho_{a_{i_1}} \cdots \varrho_{a_{i_{r-k}}} T_{A_1 \to \pi'} \delta_\pi \\
&+ \sum_{i=1}^{s-k} \varrho^*_{b_{j_1}} \cdots \varrho^*_{b_{j_{i-1}}} \varrho^*_{b_{j_{i+1}}} \cdots \varrho^*_{b_{j_{s-k}}} T_{\{a_{r+1}\} \to b_{j_i}} \varrho_{a_{i_1}} \cdots \varrho_{a_{i_{r-k}}} T_{A_1 \to \pi'} \delta_\pi \Big] \\
&+ \sum_{\substack{A_1 \sqcup \{a_{i_1}, \dots, a_{i_{r-s}}\} = \{a_1, \dots, a_r\} \\ |A_1| = s}} \sum_{\pi' \in \Gamma(\{b_1, \dots, b_s\})} \varrho_{a_{r+1}} \varrho_{a_{i_1}} \cdots \varrho_{a_{i_{r-s}}} T_{A_1 \to \pi'} \delta_\pi \\
&= \sum_{k=0}^{s} \sum_{\substack{A_1 \sqcup \{a_{i_1}, \dots, a_{i_{r+1-k}}\} = \{a_1, \dots, a_{r+1}\} \\ B_1 \sqcup \{b_{j_1}, \dots, b_{j_{s-k}}\} = \{b_1, \dots, b_s\} \\ |A_1| = |B_1| = k}} \sum_{\pi' \in \Gamma(B_1)} \varrho^*_{b_{j_1}} \cdots \varrho^*_{b_{j_{s-k}}} \varrho_{a_{i_1}} \cdots \varrho_{a_{i_{r+1-k}}} T_{A_1 \to \pi'} \delta_\pi .
\end{aligned}$$

In both cases the proof is concluded.
4. The proof of Property 4. is fully analogous to the one of Proposition 61. It is left to the reader. □

Property 3 from Lemma 144 is the analogue of Lemma 49. It allows to prove Theorem 124.

*Proof of Theorem 124.* One clearly has $\phi'_{[\![n]\!]}(H^0) = V^0$ and $V^0 \cong S^{(n)}$. Let $k \in \{2, \dots, n\}$ and $A \in \mathcal{P}([\![n]\!])$ with $|A| = k$. We define the space $W^k_A = W^k \cap \operatorname{span}\{\mathbb{1}_{\mathfrak{S}_n(\pi)} \mid \pi \in \Gamma(A)\}$. We first prove that $\phi'_{[\![n]\!]}(H_A) \subset W^k_A$. Let $F \in H_A$ and let $B \in \mathcal{P}([\![n]\!])$ with $|B| \leq k-1$. By definition $\phi'_{[\![n]\!]}(H_A) \subset \operatorname{span}\{\mathbb{1}_{\mathfrak{S}_n(\pi)} \mid \pi \in \Gamma(A)\}$. We then need to prove that $M_B \phi'_{[\![n]\!]} F = 0$. Properties 2. and 3. of Lemma 144 give

$$M_B \phi'_{[\![n]\!]} F = M_B \phi'_{A \cup B} F = \sum_{\substack{A_1 \subset A \setminus B \\ B_1 \subset B \setminus A \\ |A_1| = |B_1|}} \sum_{\pi' \in \Gamma(B_1)} \phi'_B M_{(A \cap B) \sqcup B_1} T_{A_1 \to \pi'} F.$$

The space $H^k$ being stable under translations, one has $T_{A_1 \to \pi'} F \in H^k$ for any $A_1 \subset A$ and $\pi' \in \Gamma^{|A_1|}$. Now, for any $B_1 \subset B \setminus A$, $|(A \cap B) \sqcup B_1| = |A \cap B| + |B_1| \leq |B| \leq k-1$. Hence $M_{(A \cap B) \sqcup B_1} T_{A_1 \to \pi'} F = 0$ and $M_B \phi'_{[\![n]\!]} F = 0$. One therefore has $\phi'_{[\![n]\!]}(H_A) \subset W^k_A$. In addition, for $F \in H_A$ such that $\phi'_{[\![n]\!]} F = 0$, property 2. of Lemma 144 gives $0 = M_A \phi'_{[\![n]\!]} F = F$. The operator $\phi'_{[\![n]\!]}$ is thus an injection from $H_A$ to $W^k_A$ and thus $\dim W^k_A \geq d_k$ by Theorem 51. Now, by construction $W^k = \bigoplus_{|A|=k} W^k_A$, so that

$$n! = \dim \left( V^0 \oplus \bigoplus_{k=2}^{n} \bigoplus_{|A|=k} W^k_A \right) \leq 1 + \sum_{k=2}^{n} \binom{n}{k} d_k = n!.$$

Hence all the inequalities are equalities and therefore $\phi'_{[\![n]\!]}(H^k) = W^k$. Property 4. of Lemma 144 then ensures that $W^k \cong H^k$. □

## D.2 Proofs of Section 6.3

### D.2.1 Proofs of Subsection 6.3.1

The proofs of Theorem 129 and Proposition 130 require the two following lemmas. The proof of the first one is straightforward and left to the reader.

**Lemma 145.** *For $a, b, c \in [\![n]\!]$ with $b \neq c$ one has*

$$\langle e_a, e_b \rangle = \begin{cases} -1 & \text{if } a \neq b \\ n-1 & \text{if } a = b \end{cases} \qquad \text{and} \qquad \langle e_a, x_{b \succ c} \rangle = \begin{cases} 1 & \text{if } a = b \\ -1 & \text{if } a = c \\ 0 & \text{if } a \notin \{b, c\} \end{cases}$$

**Lemma 146.** *For $a, b \in [\![n]\!]$ with $a \neq b$ and $s \in \mathbb{R}^n$ one has*

$$\sum_{1 \leq i < j \leq n} (s_i - s_j) x_{i \succ j} = \sum_{i \in [\![n]\!]} s_i e_i \qquad \text{and} \qquad f_{a,b} = n\, x_{a \succ b} + e_b - e_a.$$

*Proof.* Recalling that for any $i, j \in [\![n]\!]$ with $i \neq j$, $x_{j \succ i} = -x_{i \succ j}$, straightforward calculations give

$$\sum_{1 \leq i < j \leq n} (s_i - s_j) x_{i \succ j} = \frac{1}{2} \sum_{1 \leq i \neq j \leq n} (s_i - s_j) x_{i \succ j} = \sum_{i \in [\![n]\!]} s_i \sum_{j \neq i} x_{i \succ j} + \sum_{j \in [\![n]\!]} s_j \sum_{i \neq j} x_{j \succ i} = \sum_{i \in [\![n]\!]} s_i e_i$$

and

$$f_{a,b} = \sum_{c \notin \{a,b\}} (x_{a \succ b} + x_{b \succ c} + x_{c \succ a}) = (n-2) x_{a \succ b} + (e_b - x_{b \succ a}) - (e_a - x_{a \succ b}) = n\, x_{a \succ b} + e_b - e_a.$$

□

*Proof of Theorem 129.* We first show that the spaces $H_1^2$ and $H_2^2$ are orthogonal. Let $a, b, c \in [\![n]\!]$ with $b \neq c$. By Lemmas 145 and 146, one has

$$\langle e_a, f_{b,c} \rangle = n \langle e_a, x_{b \succ c} \rangle + \langle e_a, e_c \rangle - \langle e_a, e_b \rangle = \begin{cases} n - 1 - (n-1) = 0 & \text{if } a = b \\ -n + (n-1) + 1 = 0 & \text{if } a = c \\ 0 - 1 + 1 = 0 & \text{if } a \notin \{b, c\} \end{cases}.$$

Next we prove that $H_1^2$ and $H_2^2$ are both representations of $\mathfrak{S}_n$, or equivalently stable under translations. For $a, b \in [\![n]\!]$ with $a \neq b$ and $\tau \in \mathfrak{S}_n$ one has by definition $T_\tau x_{a \succ b} = x_{\tau(a) \succ \tau(b)}$, so that

$$T_\tau e_a = \sum_{c \neq a} x_{\tau(a) \succ \tau(c)} = \sum_{c \neq a} x_{\tau(a) \succ c} = e_{\tau(a)}$$

and

$$\begin{aligned} T_\tau f_{a,b} &= \sum_{c \notin \{a,b\}} (x_{\tau(a) \succ \tau(b)} + x_{\tau(b) \succ \tau(c)} + x_{\tau(c) \succ \tau(a)}) \\ &= \sum_{c \notin \{a,b\}} T_\tau (x_{\tau(a) \succ \tau(b)} + x_{\tau(b) \succ c} + x_{c \succ \tau(a)}) = f_{\tau(a), \tau(b)}. \end{aligned}$$

Now, Theorem 122 ensures that $H^2 \cong S^{(n-1,1)} \oplus S^{(n-2,1,1)}$ as representations of $\mathfrak{S}_n$, where $S^{(n-1,1)}$ and $S^{(n-2,1,1)}$ are both irreducible representations. Since $H_1^2 \neq \{0\}$, one then necessarily has $H_1^2 \cong S^{(n-1,1)}$ and $H_2^2 \cong S^{(n-2,1,1)}$ or $H_2^2 \cong S^{(n-1,1)}$ and $H_1^2 \cong S^{(n-2,1,1)}$. To conclude, notice that since $H_1^2 = \operatorname{span}\{e_a \mid a \in [\![n]\!]\}$, $\dim H_1^2 \leq n < \binom{n-1}{2}$ and one cannot have $H_1^2 \cong S^{(n-2,1,1)}$. Hence the other alternative is true and this concludes the proof. □

### D.2.2 Proofs of Subsection 6.3.2

*Proof of Lemma 131.* We prove the formulas in order.

(i) By definition $x_{a\succ b} = ab - ba$ so $K_{a,b} = \phi'_{[\![n]\!]}(x_{a\succ b}) = \mathbb{1}_{\mathfrak{S}_n(ab)} - \mathbb{1}_{\mathfrak{S}_n(ba)}$ and $K_{a,b}(\sigma) = \mathbb{I}\{\sigma(a) < \sigma(b)\} - \mathbb{I}\{\sigma(a) > \sigma(b)\}$.

(ii) One has

$$B_a(\sigma) = \sum_{c\neq a} K_{a,c}(\sigma) = \sum_{c\neq a} \mathbb{I}\{\sigma(a) < \sigma(c)\} - \mathbb{I}\{\sigma(a) > \sigma(c)\} = \sum_{j=\sigma(a)+1}^{n} 1 - \sum_{j=1}^{\sigma(a)-1} 1.$$

(iii) By Lemma 146 one has $C_{a,b} = \phi'_{[\![n]\!]}(n\, x_{a\succ b} + e_b - e_a) = nK_{a,b} + B_b - B_a$. Then

$$C_{a,b}(\sigma) = n\,\mathrm{sign}(\sigma(b)-\sigma(a)) + (n+1-2\sigma(b)) - (n+1-2\sigma(a)) = n\,\mathrm{sign}(\sigma(b)-\sigma(a)) - 2(\sigma(b)-\sigma(a)).$$

Equivalently,

$$C_{a,b} = \sum_{\substack{r=-n+1 \\ r\neq 0}}^{n-1} (n\,\mathrm{sign}(r) - 2r)\mathbb{1}_{\{\sigma(b)-\sigma(a)=r\}} = \sum_{r=1}^{n-1} (n-2r)\left(\mathbb{1}_{\{\sigma(b)-\sigma(a)=r\}} - \mathbb{1}_{\{\sigma(b)-\sigma(a)=-r\}}\right).$$

□

The proofs of Theorem 133 and Lemma 134 will extensively use the following lemma.

**Lemma 147** (Inner products)**.** *Let $a, b, c, d \in [\![n]\!]$ with $a \neq b$ and $c \neq d$. All the values of the inner products between $K_{i,j}$'s, $B_i$'s and $C_{i,j}$'s are given in the following tables.*

| *Condition* | $\langle B_a, K_{c,d}\rangle$ | $\langle B_a, B_c\rangle$ | $\langle B_a, C_{c,d}\rangle$ |
|---|---|---|---|
| $a = c$ | $(n+1)!/3$ | $(n-1)(n+1)!/3$ | $0$ |
| $a = d$ | $-(n+1)!/3$ | $-(n+1)!/3$ | $0$ |
| $a \notin \{c, d\}$ | $0$ | $0$ | $0$ |

| *Condition* | | $\langle K_{a,b}, K_{c,d}\rangle$ | $\langle C_{a,b}, K_{c,d}\rangle$ | $\langle C_{a,b}, C_{c,d}\rangle$ |
|---|---|---|---|---|
| $\lvert\{a,b\}\cap\{c,d\}\rvert = 2$ | $a = c, b = d$ | $n!$ | $(n-2)n!/3$ | $(n-2)n\,n!/3$ |
| $\lvert\{a,b\}\cap\{c,d\}\rvert = 2$ | $a = d, b = c$ | $-n!$ | $-(n-2)n!/3$ | $-(n-2)n\,n!/3$ |
| $\lvert\{a,b\}\cap\{c,d\}\rvert = 1$ | $a = c, b \neq d$ | $n!/3$ | $-n!/3$ | $-n\,n!/3$ |
| $\lvert\{a,b\}\cap\{c,d\}\rvert = 1$ | $a \neq c, b = d$ | $n!/3$ | $-n!/3$ | $-n\,n!/3$ |
| $\lvert\{a,b\}\cap\{c,d\}\rvert = 1$ | $a = d, b \neq c$ | $-n!/3$ | $n!/3$ | $n\,n!/3$ |
| $\lvert\{a,b\}\cap\{c,d\}\rvert = 1$ | $a \neq d, b = c$ | $-n!/3$ | $n!/3$ | $n\,n!/3$ |
| $\lvert\{a,b\}\cap\{c,d\}\rvert = 0$ | | $0$ | $0$ | $0$ |

*Proof.* First we calculate the inner product between $K_{a,b}$ and $K_{c,d}$.

$$\begin{aligned}\langle K_{a,b}, K_{c,d}\rangle &= \left\langle \mathbb{1}_{\mathfrak{S}_n(ab)} - \mathbb{1}_{\mathfrak{S}_n(ba)}, \mathbb{1}_{\mathfrak{S}_n(cd)} - \mathbb{1}_{\mathfrak{S}_n(dc)} \right\rangle \\ &= |\mathfrak{S}_n(ab) \cap \mathfrak{S}_n(cd)| - |\mathfrak{S}_n(ab) \cap \mathfrak{S}_n(dc)| - |\mathfrak{S}_n(ba) \cap \mathfrak{S}_n(cd)| + |\mathfrak{S}_n(ba) \cap \mathfrak{S}_n(dc)|\end{aligned}$$

Now one has

$$|\mathfrak{S}_n(ab) \cap \mathfrak{S}_n(cd)| = \begin{cases} \text{(if } a = c, b = d\text{)} & |\mathfrak{S}_n(ab)| = n!/2 \\ \text{(if } a = d, b = c\text{)} & |\emptyset| = 0 \\ \text{(if } a = c, b \neq d\text{)} & |\mathfrak{S}_n(a \succ b, d)| = n!/6 + n!/6 = n!/3 \\ \text{(if } a \neq c, b = d\text{)} & |\mathfrak{S}_n(a, c \succ d)| = n!/6 + n!/6 = n!/3 \\ \text{(if } a = d, b \neq c\text{)} & |\mathfrak{S}_n(c \succ a \succ b)| = n!/6 \\ \text{(if } a \neq d, b = c\text{)} & |\mathfrak{S}_n(a \succ b \succ d)| = n!/6 \\ \text{(if } \{a,b\} \cap \{c,d\} = \emptyset\text{)} & |\mathfrak{S}_n(a \succ b \text{ and } c \succ d)| = n!/4 \end{cases}$$

Injecting these values provides all the results. Then we calculate

1. $\langle B_a, K_{c,d}\rangle$ using $B_a = \sum_{b \neq a} K_{a,b}$ and $\langle K_{a,b}, K_{c,d}\rangle$
2. $\langle B_a, B_c\rangle$ using $B_c = \sum_{d \neq c} K_{c,d}$ and $\langle B_a, K_{c,d}\rangle$
3. $\langle B_a, C_{c,d}\rangle$ using $C_{c,d} = nK_{c,d} + B_d - B_c$, $\langle B_a, B_c\rangle$ and $\langle B_a, K_{c,d}\rangle$
4. $\langle C_{a,b}, K_{c,d}\rangle$ using $C_{a,b} = nK_{a,b} + B_b - B_a$, $\langle K_{a,b}, K_{c,d}\rangle$ and $\langle B_a, K_{c,d}\rangle$
5. $\langle C_{a,b}, C_{c,d}\rangle$ using $C_{a,b} = nK_{a,b} + B_b - B_a$, $\langle K_{a,b}, K_{c,d}\rangle$, $\langle B_a, K_{c,d}\rangle$ and $\langle B_a, B_c\rangle$

□

*Proof of Theorem 133.* We first give another formulation of $R_2 f$ for any $f \in L(\mathfrak{S}_n)$. By Proposition 125,

$$R_2 f = (n-2)! \left(\frac{2}{n!}\right)^2 \sum_{1 \leq a \neq b \leq n} \left\langle f, \mathbb{1}_{\mathfrak{S}_n(ab)} \right\rangle \mathbb{1}_{\mathfrak{S}_n(ab)}.$$

Then, using the fact that $2\mathbb{1}_{\mathfrak{S}_n(ab)} = \mathbb{1}_{\mathfrak{S}_n} + K_{a,b}$ for any $a, b \in [\![n]\!]$ with $a \neq b$,

$$\begin{aligned}\sum_{1 \leq a \neq b \leq n} \left\langle f, \mathbb{1}_{\mathfrak{S}_n(ab)} \right\rangle \mathbb{1}_{\mathfrak{S}_n(ab)} &= \sum_{1 \leq a < b \leq n} \left\langle f, \mathbb{1}_{\mathfrak{S}_n(ab)} \right\rangle \mathbb{1}_{\mathfrak{S}_n(ab)} + \left\langle f, \mathbb{1}_{\mathfrak{S}_n(ba)} \right\rangle \mathbb{1}_{\mathfrak{S}_n(ba)} \\ &= \frac{1}{2} \sum_{1 \leq a < b \leq n} \left\langle f, \mathbb{1}_{\mathfrak{S}_n(ab)} \right\rangle (\mathbb{1}_{\mathfrak{S}_n} + K_{a,b}) + \left\langle f, \mathbb{1}_{\mathfrak{S}_n(ba)} \right\rangle (\mathbb{1}_{\mathfrak{S}_n} - K_{a,b}) \\ &= \frac{1}{2} \sum_{1 \leq a < b \leq n} \langle f, \mathbb{1}_{\mathfrak{S}_n} \rangle \mathbb{1}_{\mathfrak{S}_n} + \langle f, K_{a,b} \rangle K_{a,b}.\end{aligned}$$

This gives the following formula

$$R_2 f = \frac{1}{n!} \langle f, \mathbb{1}_{\mathfrak{S}_n} \rangle \mathbb{1}_{\mathfrak{S}_n} + \frac{1}{2\binom{n}{2} n!} \sum_{1 \leq a \neq b \leq n} \langle f, K_{a,b} \rangle K_{a,b}. \tag{D.4}$$

Injecting Lemma 147 in Formula (D.4) then gives, for any $a, b \in [\![n]\!]$ with $a \neq b$, (notice that as $\langle K_{a,b}, \mathbb{1}_{\mathfrak{S}_n} \rangle = |\mathfrak{S}_n(ab)| - |\mathfrak{S}_n(ba)| = 0$, $\langle B_a, \mathbb{1}_{\mathfrak{S}_n} \rangle = \langle C_{a,b}, \mathbb{1}_{\mathfrak{S}_n} \rangle = 0$):

$$R_2 \mathbb{1}_{\mathfrak{S}_n} = \mathbb{1}_{\mathfrak{S}_n}, \qquad R_2 B_a = \frac{(n+1)}{3\binom{n}{2}} B_a \qquad \text{and} \qquad R_2 C_{a,b} = \frac{1}{3\binom{n}{2}} C_{c,d}.$$

This shows that $V^0$, $\mathscr{B}_n$ and $\mathscr{C}_n$ are included in the eigenspaces of $R_2$ for the respective eigenvalues 1, $(n+1)/(3\binom{n}{2})$ and $1/(3\binom{n}{2})$. Now, $\dim V^0 + \dim \mathscr{B}_n + \dim \mathscr{C}_n = 1 + (n-1) + \binom{n-1}{2} = 1 + \binom{n}{2}$ and $\dim \ker R_2 = n! - \binom{n}{2} - 1$ by Theorem 127. Hence all the inclusions are equalities and the proof is concluded. □

*Proof of Lemma 134.* For any subspace $V$ of $L(\mathfrak{S}_n)$, the orthogonal projection of $p$ on $V$ is the unique element $p'$ of $V$ such that $\langle p - p', q \rangle = 0$ or equivalently that $\langle p, q \rangle = \langle p', q \rangle$ for all $q \in V$. As $\mathscr{B}_n = \operatorname{span}(B_a)_{1 \leq a \leq n}$ and $\mathscr{C}_n = \operatorname{span}(C_{a,b})_{1 \leq a \neq b \leq n}$, it is sufficient to show that $\langle p, B_a \rangle = \langle p_{\mathscr{B}_n}, B_a \rangle$ and $\langle p, C_{a,b} \rangle = \langle p_{\mathscr{C}_n}, C_{a,b} \rangle$ for all $a, b \in [\![n]\!]$ with $a \neq b$. This is done using Lemma 147. □

### D.2.3 Proofs of Subsection 6.3.3

*Proof of Proposition 136.* We begin with the left inequality. For $\sigma \in \mathfrak{S}_n$ one has by Lemma 134

$$|p_{\mathscr{C}_n}(\sigma)| = \frac{3}{n^2 n!} \left| \sum_{\{a,b\} \subset [\![n]\!]} \langle p, C_{a,b} \rangle \, C_{a,b}(\sigma) \right| \leq \frac{3(n-2)}{n^2 n!} \sum_{\{a,b\} \subset [\![n]\!]} |\langle p, C_{a,b} \rangle|,$$

because $\|C_{a,b}\|_\infty = (n-2)$ for any $a, b \in [\![n]\!]$ with $a \neq b$ by Lemma 131. Now, by definition of $C_{a,b}$,

$$|\langle p, C_{a,b} \rangle| = \left| \left\langle p, \sum_{c \notin \{a,b\}} K_{a,b} + K_{b,c} + K_{c,a} \right\rangle \right| \leq \sum_{c \notin \{a,b\}} \mathrm{Cyc}_{\{a,b,c\}}(p).$$

Combining the two gives the left inequality. For the right inequality, let $a, b, c \in [\![n]\!]$ be distinct elements. First observe that

$$C_{a,b} + C_{b,c} + C_{c,a} = n(K_{a,b} + K_{b,c} + K_{c,a})$$

by Lemma 131. One therefore has

$$\mathrm{Cyc}_{\{a,b,c\}}(p) = |\langle p, K_{a,b} + K_{b,c} + K_{c,a} \rangle| \leq \frac{1}{n} \Big( |\langle p, C_{a,b} \rangle| + |\langle p, C_{b,c} \rangle| + |\langle p, C_{c,b} \rangle| \Big).$$

Hence

$$\sum_{\{a,b,c\} \subset [\![n]\!]} \mathrm{Cyc}_{\{a,b,c\}}(p) \leq \frac{3(n-2)}{n} \sum_{\{a,b\} \subset [\![n]\!]} |\langle p, C_{a,b} \rangle| \leq \frac{3(n-2)}{n} \binom{n}{2} \left( \sum_{\{a,b\} \subset [\![n]\!]} \langle p, C_{a,b} \rangle^2 \right)^{1/2},$$

where the last part comes from the Cauchy-Schwarz inequality. On the other hand one has

$$\|p_{\mathscr{C}_n}\|_2^2 = \langle p_{\mathscr{C}_n}, p_{\mathscr{C}_n} \rangle = \frac{3}{n^2 n!} \sum_{\{a,b\} \subset [\![n]\!]} \langle p, C_{a,b} \rangle \langle p_{\mathscr{C}_n}, C_{a,b} \rangle = \frac{3}{n^2 n!} \sum_{\{a,b\} \subset [\![n]\!]} \langle p, C_{a,b} \rangle^2,$$

where the last equality comes from the fact that $p_{\mathscr{C}_n}$ is the orthogonal projection of $p$ onto $\mathscr{C}_n = \operatorname{span}(C_{a,b})_{1 \leq a < b \leq n}$ and therefore satisfies $\langle p_{\mathscr{C}_n}, C_{a,b} \rangle = \langle p, C_{a,b} \rangle$ for all $a, b \in [\![n]\!]$ with $a \neq b$. Combining the two concludes the proof. □

# Appendix E

# Condensé en français

## Introduction

On s'intéresse aux problèmes de *ranking* sur un nombre fini (mais potentiellement grand) d'éléments. Le cadre naturel de modélisation est donc le suivant :
On dispose de $n$ objets numérotés arbitrairement de 1 à $n$. Un *ranking* de ces $n$ objets est une façon de les ordonner, ce que l'on peut modéliser par une permutation $\sigma$ de $\{1, \dots, n\}$, où

- $\sigma(i)$ est le rang de l'objet numéroté $i$
- $\sigma^{-1}(i)$ est le numéro de l'objet classé à la $i$ème place

Dans les applications visées, on est naturellement amené à manipuler des probabilités sur l'ensemble des permutations de $\{1, \dots, n\}$, noté $\mathfrak{S}_n$. Le principal problème est qu'une telle probabilité nécessite $n! - 1$ paramètres pour être caractérisée et donc stockée. Or, cette quantité dépasse les capacités de stockages actuelles dès que $n$ dépasse 15. Et nous aimerions avoir $n = 1\ 000\ 000\dots$

Notre problème est donc : **Trouver une représentation efficace pour les probabilités sur le groupe symétrique**.

Cette représentation doit permettre, en plus de pouvoir représenter les probabilités sur le groupe symétrique avec peu de paramètres, de résoudre les problèmes classiques :

- raisonnement bayésien
- apprentissage supervisé
- *clustering*
- optimisation

En outre, il faut pouvoir l'utiliser pour différents types d'observations sur la probabilité :

- les valeurs pour quelques permutations
- des coefficients de fourier
- des *rankings* partiels et/ou incomplets

# E.1 Cadre général

## E.1.1 Définitions

**Algèbre de convolution** Soit $G$ un groupe fini. Un signal sur G est modélisé au sens large par une fonction de $G$ dans $\mathbb{C}$.
On note $\mathbb{C}[G] = \{f : G \to \mathbb{C}\}$.
Pour $g \in G$, on note $\delta_g$ sa fonction indicatrice.
Chaque élément de $\mathbb{C}[G]$ se décompose ainsi de manière canonique :

$$f = \sum_{g \in G} f(g)\delta_g$$

$\mathbb{C}[G]$ est donc naturellement isomorphe à l'espace vectoriel hermitien $\mathbb{C}^{|G|}$ et hérite du produit scalaire et de la norme :

$$\langle \phi, \psi \rangle = \sum_{g \in G} \phi(g)\overline{\psi(g)} \qquad \text{et} \qquad \|\phi\| = \left( \sum_{g \in G} |\phi(g)|^2 \right)^{\frac{1}{2}}$$

La base canonique $\{\delta_g, g \in G\}$ est orthonormée pour ce produit scalaire.
On définit le produit de convolution de deux éléments de $\mathbb{C}[G]$ par :

$$(\phi * \psi)(g) = \sum_{h \in G} \phi\left(gh^{-1}\right)\psi(h)$$

On a : $\forall g, h \in G, \quad \delta_g * \delta_h = \delta_{gh}$.
$(\mathbb{C}[G], +, *)$ est alors une algèbre, on l'appelle l'algèbre du groupe $G$.

**La représentation régulière gauche, ou l'action des translations** A tout $g \in G$, on peut associer la translation $T_g \in \mathcal{L}(\mathbb{C}[G])$ définie par

$$T_g f = \delta_g * f, \qquad i.e. \qquad \forall h \in G, \quad T_g f(h) = f(g^{-1}h)$$

C'est une application inversible avec $(T_g)^{-1} = T_{g^{-1}}$.
De plus, on a : $\forall g, h \in G, \quad T_g T_h = T_{gh}$.
Ainsi, $g \mapsto T_g$ est un morphisme de groupes entre $G$ et $GL(\mathbb{C}[G])$, que l'on note $\rho_{reg}$.
$(\mathbb{C}[G], \rho_{reg})$ est appelée la représentation régulière gauche du groupe $G$.

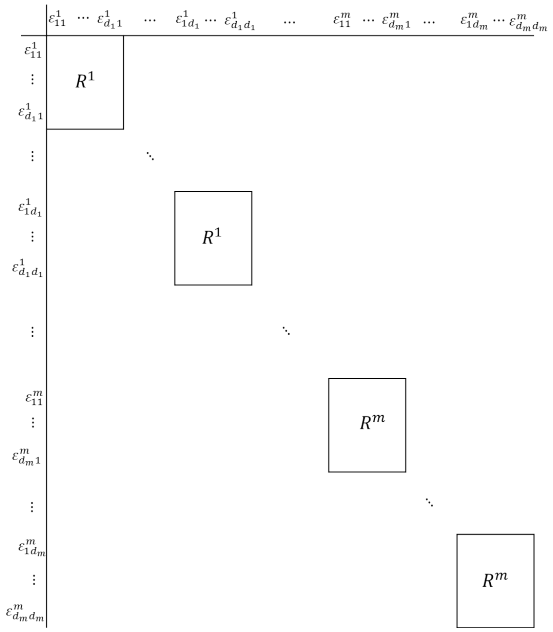
$G$ étant un groupe fini, il admet un nombre fini de représentations irréductibles inéquivalentes. Le théorème de Mashke nous dit que toute représentation (sur $\mathbb{C}$) est équivalente à une somme de ces représentations irréductibles. Dans le cas de la représentation régulière, on peut être plus précis. On note $\rho^1, ..., \rho^m$ un représentant de chaque représentation irréductible, où $\rho^k$ agit sur un espace de dimension $d_k$. On a alors :

$$\rho_{reg} \simeq \bigoplus_{k=1}^{m} d_k \rho^k$$

De plus, on sait qu'il est possible de représenter chaque $\rho^k$ par une matrice unitaire $R^k \in GL_{d_k}(\mathbb{C})$. On pose alors :

$$\text{pour} \quad k \in \{1, ..., m\}, \quad i, j \in \{1, ..., d_k\}, \quad \varepsilon_{ij}^k : g \mapsto \sqrt{\frac{d_k}{|G|}}\overline{R_{ij}^k(g)}$$

Et d'après le lemme de Schur (cf Diaconis), la famille $\left\{\varepsilon_{ij}^k \,|\, k \in \{1,...,m\}\,,\ i,j \in \{1,...,d_k\}\right\}$ est une base orthonormée de $\mathbb{C}[G]$. De plus, pour tout $g \in G$, la matrice de $\rho_{reg}(g)$ dans cette base est diagonale par blocs avec :




**Les projection sur les sous-espaces isotypiques et la transformée de Fourier** Pour $k \in \{1,...,m\}$ et $j \in \{1,...,d_k\}$, on pose :

$$V_j^k = Vect\left(\varepsilon_{1j}^k, \varepsilon_{2j}^k, ..., \varepsilon_{d_k j}^k\right) \qquad \text{et} \qquad V^k = \bigoplus_{j=1}^{d_k} V_j^k$$

Comme nous l'avons vu, chaque $V_j^k$ est stable sous l'action de $\{\rho_{reg}(g), g \in G\}$, et donc $V^k$ aussi. On l'appelle le sous-espace isotypique associé à la représentation irréductible $\rho^k$.
Pour $f \in \mathbb{C}[G]$, on note $\pi^k f$ sa projection (orthogonale) sur $V^k$.
$\left\{\varepsilon_{ij}^k \,|\, k \in \{1,...,m\}\,,\ i,j \in \{1,...,d_k\}\right\}$ étant une base orthonormée, on a les formules :

$$\langle \phi, \psi \rangle = \sum_{k=1}^{m} \sum_{1 \leqslant i,j \leqslant d_k} \left\langle \phi, \varepsilon_{ij}^k \right\rangle \overline{\left\langle \psi, \varepsilon_{ij}^k \right\rangle}$$

$$\|f\| = \left( \sum_{k=1}^{m} \sum_{1 \leqslant i,j \leqslant d_k} |\left\langle f, \varepsilon_{ij}^k \right\rangle|^2 \right)^{\frac{1}{2}}$$

$$f = \sum_{k=1}^{m} \sum_{1 \leqslant i,j \leqslant d_k} \left\langle f, \varepsilon_{ij}^k \right\rangle \varepsilon_{ij}^k$$

$$\pi^k f = \sum_{1\leqslant i,j\leqslant d_k} \left\langle f, \varepsilon_{ij}^k \right\rangle \varepsilon_{ij}^k$$

La projection de $f$ sur $V^k$ est ainsi donnée par les coefficients $\left\{\left\langle f, \varepsilon_{ij}^k \right\rangle \mid 1 \leqslant i,j \leqslant d_k\right\}$. Il est donc naturel de considérer la matrice $\left(\left\langle f, \varepsilon_{ij}^k \right\rangle\right)_{1\leqslant i,j\leqslant d_k} \in \mathcal{M}_{d_k}(\mathbb{C})$. En fait, on pose :

$$\widehat{f}\left(\rho^k\right) = \sqrt{\frac{|G|}{d_k}} \left(\left\langle f, \varepsilon_{ij}^k \right\rangle\right)_{1\leqslant i,j\leqslant d_k}$$

Ce qui donne :

$$\widehat{f}\left(\rho^k\right) = \sum_{g\in G} f(g) R^k(g)$$

On appelle $\widehat{f}$ la transformée de Fourier de $f$. On la note aussi $\mathcal{F}f$.
On munit $\mathcal{M}_{d_k}(\mathbb{C})$ du produit scalaire $\langle A, B\rangle = Tr\left(A^* B\right)$.
Il s'agit du produit scalaire coordonnée par coordonnée, donc $\widehat{f}\left(\rho^k\right)$ hérite pour ce produit scalaire de toutes les propriétés de $\pi^k f$ en tant que projection orthogonale. On obtient ainsi les formules dites respectivement de Plancherel, d'isométrie, et d'inversion :

$$\langle \phi, \psi\rangle = \frac{1}{|G|} \sum_{k=1}^{m} d_k \left\langle \widehat{\phi}\left(\rho^k\right), \widehat{\psi}\left(\rho^k\right)\right\rangle$$

$$\|f\|^2 = \frac{1}{|G|} \sum_{k=1}^{m} d_k \|\widehat{f}\left(\rho^k\right)\|^2$$

$$f(g) = \frac{1}{|G|} \sum_{k=1}^{m} d_k \left\langle R^k(g)\widehat{f}\left(\rho^k\right)\right\rangle = \frac{1}{|G|} \sum_{k=1}^{m} d_k Tr\left[R^k(g^{-1})\widehat{f}\left(\rho^k\right)\right]$$

Pour la dernière formule, on a utilisé le fait que $R^k$ est unitaire. En particulier, on a aussi :

$$\pi^k f(g) = \frac{d_k}{|G|} Tr\left[R^k(g^{-1})\widehat{f}\left(\rho^k\right)\right]$$

Enfin, comme pour les fonctions réelles, la transformée de Fourier transforme le produit de convolution en produit (ici matriciel) :

$$\widehat{\phi * \psi}\left(\rho^k\right) = \widehat{\phi}\left(\rho^k\right)\widehat{\psi}\left(\rho^k\right)$$

**Graphes de Cayley et Laplacien** Dans le cas des fonctions réelles, la transformée de Fourier interagit agréablement avec la dérivation, ou aussi, avec le laplacien. D'ailleurs, la décomposition de Fourier dans $L^2([0,1])$ peut s'obtenir par l'étude spectrale du laplacien, vu comme opérateur auto-adjoint compact et positif. Dans le cas d'un groupe fini, il est possible de faire un lien entre la transformée de Fourier définie précédemment et le laplacien sur certains graphes, dits de Cayley.

Soit $S \subset G$. On note $\Gamma(G,S)$ le graphe qui a pour ensemble de noeuds $G$, et où $g$ est relié à $h$ (que l'on note $g \sim h$) si $hg^{-1} \in S$. Les propriétés suivantes sont immédiates :

- $\Gamma(G,S)$ est sans boucle ssi $id \notin S$
- $\Gamma(G,S)$ est non orienté ssi $S^{-1} = S$

- $\Gamma(G,S)$ est connexe ssi $\langle S\rangle = G$

Si $S$ vérifie ces 3 hypothèses, on dit que $\Gamma(G,S)$ est un graphe de Cayley. Dans ce cas, on a en particulier que chaque noeud a exactement $|S|$ voisins (on dit que le graphe est $|S|$-régulier). Si de plus, $S$ est invariante par conjugaison, on dit que $\Gamma(G,S)$ est quasi-abélien.

On suppose dans la suite que $\Gamma(G,S)$ est de Cayley. On définit sa matrice d'adjacence comme la matrice $\mathbf{A}(G,S)$ telle que $\mathbf{A}(G,S)_{g,h} = \mathbb{1}_{\{g\sim h\}}$. C'est une matrice $|G|\times|G|$, donc on peut la voir comme un opérateur sur $\mathbb{C}[G]$, avec :

$$\mathbf{A}(G,S)\,\delta_g = \sum_{h\sim g}\delta_h = \sum_{s\in S}\delta_{sg} = \sum_{s\in S}\rho_{reg}(s)\delta_g$$

Ainsi,

$$\mathbf{A}(G,S) = \sum_{s\in S}\rho_{reg}(s)$$

On peut aussi écrire, pour $f\in\mathbb{C}[G]$ :

$$(\mathbf{A}(G,S)\,f)(g) = \sum_{h\sim g} f(h)$$

On définit le laplacien du graphe $\Gamma(G,S)$ par $\mathbf{L}(G,S) = |S|I_{|G|} - \mathbf{A}(G,S)$. En tant qu'opérateur :

$$\mathbf{L}(G,S)\,f(g) = \sum_{h\,\mathrm{tq}\,h\sim g}(f(g)-f(h))$$

Par définition, les valeurs propres et vecteurs propres du laplacien sont directement reliés à ceux de la matrice d'adjacence. Ces deux matrices sont symétriques réelles, donc diagonalisable dans $\mathbb{R}$ avec des valeurs propres réelles. De plus, on a :

$$\langle f, \mathbf{L}(G,S)\,f\rangle = \frac{1}{2}\sum_{(g,h)\,\mathrm{tq}\,g\sim h}|f(g)-f(h)|^2$$

Donc $\mathbf{L}(G,S)$ est un opérateur positif, et ses valeurs propres sont positives.

**Lien entre les valeurs propres du laplacien et la transformée de Fourier** On a vu que $\mathbf{A}(G,S) = \sum_{s\in S}\rho_{reg}(s)$, donc elle se décompose simplement dans la base $\left\{\varepsilon_{ij}^k\right\}$ et ses valeurs propres et vecteurs propres sont donc directement reliés à ceux des $\widehat{\mathbb{1}_S}\left(\rho^k\right)$. Ainsi, l'étude spectrale du laplacien revient à l'étude spectrale des éléments matriciels de la transformée de Fourier d'une fonction de $G$ dans $\mathbb{C}$.

Dans le cas où $\Gamma(G,S)$ est quasi-abélien, $S$ est invariante par conjugaison, donc $\mathbb{1}_S$ est une fonction centrale (constante sur les classes d'équivalence). On sait alors que c'est une combinaison linéaire des caractères des représentations irréductibles, et qu'elle agit sur chaque espace isotypique comme une homothétie. Plus précisément, on a que pour tout graphe quasi-abélien, $\varepsilon_{ij}^k$ est un vecteur propre de $\mathbf{A}(G,S)$ de valeur propre : $\Lambda_k = \frac{1}{d_k}\sum_{s\in S}\chi^k(s)$ (où $\chi^k = Tr\left(\rho^k\right)$ est le caractère de la représentation irréductible $\rho^k$).

Dans le cas général, on sait que $\widehat{\mathbb{1}_S}\left(\rho^k\right)$ est diagonalisable (en tant que restriction de $\mathbf{A}(G,S)$ à $V^k$).
On note $\lambda_1^k, \lambda_2^k, ..., \lambda_{d_k}^k$ ses valeurs propres et $u_1^k, u_2^k, ..., u_{d_k}^k$ des vecteurs propres associés.
Alors : $\lambda_i^k$ est valeur propre de $\mathbf{A}(G,S)$ avec multiplicité $d_k$ et vecteurs propres

$$\phi_{ij}^k = \sum_{l=1}^{d_k} \left(u_i^k\right)_l \varepsilon_{lj}^k \qquad \text{pour} \qquad j = 1, ..., d_k$$

Malheureusement, il n'existe pas *a priori* de formule permettant de calculer les $\lambda_i^k$ ou les $u_i^k$. Une piste pour les étudier est peut-être la suivante. On prend une fonction $f \in \mathbb{C}[G]$ et une représentation irréductible $\rho$, qui agit sur un espace de dimension $d$. On note $\lambda_1, \lambda_2, ..., \lambda_d$ les valeurs propres de $\widehat{f}(\rho)$. On a alors :

$$\begin{aligned}
\sum_{i=1}^{d} \lambda_i^p &= Tr\left[(\mathcal{F}f(\rho))^p\right] \\
&= Tr\left[\mathcal{F}(f * \cdots * f)(\rho)\right] \\
&= Tr\left[\sum_{g\in G} (f * \cdots * f)(g)\rho(g)\right] \\
&= \sum_{g\in G} (f * \cdots * f)(g)\chi(g) \\
&= \left\langle \overline{f * \cdots * f}, \chi \right\rangle
\end{aligned}$$

Ce qui donne le système :

$$\begin{cases}
\lambda_1 + \lambda_2 + \ldots + \lambda_d = \left\langle \overline{f}, \chi \right\rangle \\
\lambda_1^2 + \lambda_2^2 + \ldots + \lambda_d^2 = \left\langle \overline{f * f}, \chi \right\rangle \\
\qquad\qquad\qquad\qquad \vdots \\
\lambda_1^d + \lambda_2^d + \ldots + \lambda_d^d = \left\langle \overline{f * \cdots * f}, \chi \right\rangle
\end{cases}$$

### E.1.2 Les représentations irréductibles du groupe symétrique

Soit $n \in \mathbb{N}^*$

**Lien avec les partitions d'entiers** Une partition de $n$ est un $r$-uplet d'entiers $\lambda = (\lambda_1, \ldots, \lambda_r)$ tels que :

$$\sum_{i=1}^{r} \lambda_i = n \qquad \text{et} \qquad \lambda_1 \geqslant \ldots \geqslant \lambda_r \geqslant 1$$

On note $\lambda \vdash n$.
Une permutation $\sigma \in \mathfrak{S}_n$ se décompose de manière unique en produit de $r$ cycles à supports disjoints (on prend en compte les points fixes). Si on les ordonne par taille de support décroissante, le r-uplet des tailles des cycles est alors une partition de $n$. On l'appelle la structure de cycles de $\sigma$.
On peut alors montrer que deux permutations $\sigma_1, \sigma_2 \in \mathfrak{S}_n$ sont conjuguées ssi elles ont la même structure de cycles.

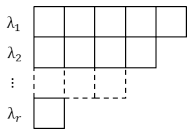
Ainsi, l'ensemble des classes de conjugaison de $\mathfrak{S}_n$ est en bijection avec l'ensemble des partitions de $n$ $\{\lambda \vdash n\}$. On note $p(n)$ son cardinal.

Un résultat classique de la théorie des représentations de groupes dit que le nombre de représentations irréductibles (ou pour être plus précis de classes d'équivalence de représentations irréductibles) d'un groupe $G$ est égal au nombre de ses classes de conjugaison.
Ainsi, $\mathfrak{S}_n$ admet $p(n)$ représentations irréductibles et on peut les indexer par les partitions de $n$. On note $\mathcal{R}_n$ l'ensemble des représentations irréductibles de $\mathfrak{S}_n$.

L'objectif est double :

- Décrire les représentations irréductibles
- Les ordonner

**Partitions d'entiers et tableaux de Young** Une partition $\lambda$ de $n$ se représente par un diagramme de Young :




Comme $\lambda$ est une partition de $n$, le diagramme a exactement $n$ cases.

Un **tableau de Young** de forme $\lambda$, encore appelé $\lambda$-tableau, est le diagramme de Young de $\lambda$ rempli avec les nombres 1, 2, ..., $n$. On note $\mathcal{T}(\lambda)$ l'ensemble des $\lambda$-tableaux. On a clairement $|\mathcal{T}(\lambda)| = n!$.
Un **tableau de Young standard** est un tableau de Young dont les entrées sont (strictement) croissantes au sein des lignes (de gauche à droite) et des colonnes (de haut en bas). On note $\mathcal{TS}(\lambda)$ l'ensemble des $\lambda$-tableaux standards.
Un **tableau de Young semistandard** est un diagramme de Young rempli avec des entiers $\geqslant 1$ de manière croissante au sein d'une ligne et strictement croissante au sein d'une colonne.
Un **tableau de Young semistandard de type** $\mu = (\mu_1, \mu_2, ..., \mu_s)$ (de forme $\lambda$) est un tableau semistandard (de forme $\lambda$) qui contient $\mu_i$ fois le nombre $i$, pour $i = 1, ..., s$.

Pour deux partitions de $n$ $\lambda = (\lambda_1, \ldots, \lambda_r)$ et $\mu = (\mu_1, \mu_2, ..., \mu_s)$, on définit la relation d'ordre :

$$\lambda \trianglerighteq \mu \qquad \text{si} \qquad \forall j \in \{1, ..., \min(r, s)\}\,, \quad \sum_{i=1}^{j} \lambda_i \leqslant \sum_{i=1}^{j} \mu_i$$

(Attention la convention n'est pas prise ici dans le même sens que dans Diaconis).
Elle définit un ordre partiel sur $\{\lambda \vdash n\}$.
On note $\lambda \triangleright \mu$ si $\lambda \trianglerighteq \mu$ et $\lambda \neq \mu$.

On appelle nombre de Kostka-$\lambda\mu$, noté $K_{\lambda\mu}$, le nombre de tableaux semistandards de forme $\lambda$ et de type $\mu$. On a alors :

$$K_{\lambda\mu} \neq 0 \;\Leftrightarrow\; \lambda \trianglelefteq \mu \qquad \text{et} \qquad K_{\lambda\lambda} = 1$$

**Action de $\mathfrak{S}_n$ sur les tableaux de Young** $\mathfrak{S}_n$ agit naturellement sur $\{1,...,n\}$ par $\sigma \cdot i = \sigma(i)$. Cette action s'étend canoniquement aux parties de $\{1,...,n\}$ à $k$ éléments par :

$$\sigma \cdot \{i_1,...,i_k\} = \{\sigma(i_1),...,\sigma(i_k)\}$$

ou encore aux $k$-uplets par :

$$\sigma \cdot (i_1,...,i_k) = (\sigma(i_1),...,\sigma(i_k))$$

On peut en fait généraliser l'approche.

On fixe $\lambda \vdash n$.
On dit que deux $\lambda$-tableaux $t_1, t_2$ sont équivalents si chaque ligne de $t_1$ contient les mêmes nombres que la ligne de $t_2$ correspondante.
Une classe d'équivalence pour cette relation s'appelle un $\lambda$-tabloïde. On note $T(\lambda)$ l'ensemble des $\lambda$-tabloïdes. Un élément $\bar{t} \in T(\lambda)$ peut se voir comme un $\lambda$-tableau dont les entrées des lignes ne sont pas ordonnées.
*On voit facilement que* $|T(\lambda)| = \frac{n!}{\lambda_1!...\lambda_r!}$

Dans le cas de certaines partitions $\lambda$, on peut interpréter les $\bar{t} \in T(\lambda)$ de manière simple :

- Si $\lambda = (n-k,k)$, $T(\lambda)$ est en bijection avec les parties de $\{1,...,n\}$ à $k$ éléments en associant à $\bar{t}$ la partie constituée des éléments de sa deuxième ligne.
- Si $\lambda = (n-k,1^k)$, $T(\lambda)$ est en bijection avec les $k$-uplets de $\{1,...,n\}$ à $k$ éléments en associant à $\bar{t}$ le $k$-uplet formé des éléments de chacune de ses lignes excepté la première.

Dans le cas général, $T(\lambda)$ est en bijection avec l'ensemble des partitions de $\{1,...,n\}$ en $r$ parties de cardinaux $\lambda_1, \lambda_2, \dots \lambda_r$.

Maintenant, l'action de $\mathfrak{S}_n$ sur $\{1,...,n\}$ s'étend aussi naturellement sur $T(\lambda)$ (c'est donc une généralisation de l'approche précédente).
On note $M^\lambda = \{f : T(\lambda) \to \mathbb{C}\}$.
Il est engendré par les fonctions indicatrices de chaque $\lambda$-tabloïde $\delta_{\bar{t}}$.
On considère alors l'action régulière de $\mathfrak{S}_n$ sur $T(\lambda)$, notée $\rho_\lambda$ :

$$\rho_\lambda(\sigma)\delta_{\bar{t}} = \delta_{\sigma\cdot\bar{t}} \qquad i.e. \qquad (\rho_\lambda(\sigma)f)\left(\bar{t}\right) = f\left(\sigma^{-1}\cdot\bar{t}\right)$$

**Les représentations irréductibles de $\mathfrak{S}_n$ : les modules de Specht** Soit $\lambda \vdash n$ et $t \in \mathcal{T}(\lambda)$ un $\lambda$-tableau.
On note $C_t$ le sous-groupe de $\mathfrak{S}_n$ constitué des permutations qui laissent stable les colonnes de $t$ (en autorisant un réordonnement).
On définit le polytabloïde associé à $t$ comme l'élément $e_t \in M^\lambda$ :

$$e_t = \sum_{\sigma\in C_t} \varepsilon(\sigma)\delta_{\sigma\cdot\bar{t}} = \left(\sum_{\sigma\in C_t} \varepsilon(\sigma)\rho_\lambda(\sigma)\right)(\delta_{\bar{t}}) = \left(\sum_{\sigma\in\mathfrak{S}_n} (\varepsilon \mathbb{1}_{C_t})(\sigma)\rho_\lambda(\sigma)\right)(\delta_{\bar{t}})$$

On définit le module de Specht $S^\lambda = Vect\{e_t \mid t \in \mathcal{T}(\lambda)\}$ .
Alors, un des résultats principaux de la théorie des représentations du groupe symétrique assure que les $S^\lambda$ sont des représentations irréductibles inéquivalentes de $\mathfrak{S}_n$.

Ainsi, on a réussi à décrire toutes les représentations irréductibles de $\mathfrak{S}_n$ en les indexant automatiquement par les partitions de $n$. On peut décrire les espaces $S^\lambda$ un peu plus précisément : $\{e_t \mid t \in \mathcal{TS}(\lambda)\}$ est une base de $S^\lambda$. Ainsi :

$$\dim S^\lambda = |\mathcal{TS}(\lambda)|$$

Il se trouve qu'on dispose de deux formules pour calculer ce nombre. La formule déterminantale :

$$\dim S^\lambda = n! \det\left(\frac{1}{(\lambda_i - i + j)!}\right)_{1 \leqslant i,j \leqslant r} \qquad \text{avec} \qquad \frac{1}{k!} = 0 \quad \text{si} \quad k < 0$$

Et la formule des équerres. Pour cela, on définit, pour $u$ une case du $\lambda$-diagramme (on note $u \in \lambda$), son équerre comme l'ensemble des cases à sa droite et en-dessous (elle y compris). Le nombre de cases dans l'équerre est appelé la longueur d'équerre de $u$, on la note $h_\lambda(u)$. On a alors :

$$\dim S^\lambda = \frac{n!}{\prod_{u \in \lambda} h_\lambda(u)}$$

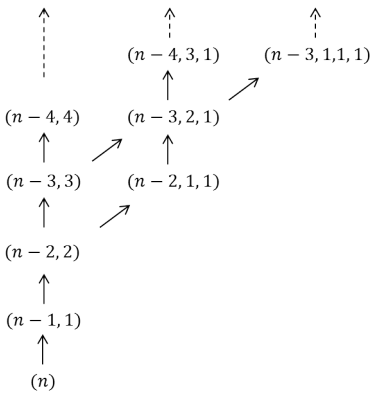
**Ordonner les représentations irréductibles** Jusque-là, nous avons réussi à décrire les représentations irréductibles de $\mathfrak{S}_n$ en les indexant par les partitions de $n$. Par ailleurs, nous avons définit un ordre partiel sur ces partitions de $n$. Nous allons donc voir comment on peut interpréter l'ordre sur les espaces $S^\lambda$. Nous allons nous appuyer sur un théorème, appelé la règle de Young :

$$M^\mu \simeq \bigoplus_{\lambda \vdash n} K_{\lambda\mu} S^\lambda$$

Or, nous avons vu que $K_{\lambda\mu} \neq 0 \;\Leftrightarrow\; \lambda \trianglelefteq \mu$ et $K_{\lambda\lambda} = 1$. On peut donc réécrire la **règle de Young** :

$$M^\mu \simeq S^\mu \oplus \bigoplus_{\lambda \lhd \mu} K_{\lambda\mu} S^\lambda$$

Voyons comment interpréter la formule pour les petites fréquences. On a le diagramme :



On a donc :

$$\begin{aligned}
M^{(n)} &= S^{(n)} \\
M^{(n-1,1)} &\simeq S^{(n)} \oplus S^{(n-1,1)} \\
M^{(n-2,2)} &\simeq S^{(n)} \oplus S^{(n-1,1)} \oplus S^{(n-2,2)} \\
M^{(n-3,3)} &\simeq S^{(n)} \oplus S^{(n-1,1)} \oplus S^{(n-2,2)} \oplus S^{(n-3,3)} \\
M^{(n-2,1,1)} &\simeq S^{(n)} \oplus 2S^{(n-1,1)} \oplus S^{(n-2,2)} \oplus S^{(n-2,1,1)} \\
M^{(n-4,4)} &\simeq S^{(n)} \oplus S^{(n-1,1)} \oplus S^{(n-2,2)} \oplus S^{(n-3,3)} \oplus S^{(n-4,4)} \\
M^{(n-3,2,1)} &\simeq S^{(n)} \oplus 2S^{(n-1,1)} \oplus 2S^{(n-2,2)} \oplus S^{(n-2,1,1)} \oplus S^{(n-3,2,1)}
\end{aligned}$$

On peut donc écrire :

$$\begin{aligned}
M^{(n)} &= S^{(n)} \\
M^{(n-1,1)} &\simeq M^{(n)} \oplus S^{(n-1,1)} \\
M^{(n-2,2)} &\simeq M^{(n-1,1)} \oplus S^{(n-2,2)} \\
M^{(n-3,3)} &\simeq M^{(n-2,2)} \oplus S^{(n-3,3)} \\
M^{(n-2,1,1)} &\simeq M^{(n-2,2)} \oplus S^{(n-1,1)} \oplus S^{(n-2,1,1)} \\
M^{(n-4,4)} &\simeq M^{(n-3,3)} \oplus S^{(n-4,4)} \\
M^{(n-3,2,1)} &\simeq M^{(n-2,1,1)} \oplus S^{(n-2,2)} \oplus S^{(n-3,2,1)}
\end{aligned}$$

Ainsi, on obtient le même diagramme pour les inclusions des espaces $M^\lambda$ que pour la relation d'ordre des partitions de $n$. De plus, la "différence" entre deux espaces $M^\lambda$ comparables s'exprime en fonctions d'espaces $S^\lambda$.

### E.1.3 Etude des distributions de probabilité sur $\mathfrak{S}_n$

**Décomposition en marginales** On a défini précédemment l'action régulière de $\mathfrak{S}_n$ sur $M^\lambda$ par :

$$\rho_\lambda(\sigma)\delta_{\overline{t}} = \delta_{\sigma\cdot\overline{t}} \qquad i.e. \qquad (\rho_\lambda(\sigma)f)\left(\overline{t}\right) = f\left(\sigma^{-1}\cdot\overline{t}\right)$$

En notant $R^\lambda(\sigma)$ la matrice de $\rho_\lambda(\sigma)$ dans la base $\left\{\delta_{\overline{t}} \mid \overline{t} \in T(\lambda)\right\}$, on voit que c'est une matrice de permutation de $\mathcal{M}_{|T(\lambda)|}(\mathbb{R})$, avec :

$$R^\lambda(\sigma)_{\overline{u},\overline{t}} = \mathbb{1}_{\left\{\sigma\cdot\overline{t}=\overline{u}\right\}}$$

Soit maintenant $f \in \mathbb{C}[\mathfrak{S}_n]$ une distribution de probabilité et $\mathbb{P}$ sa probabilité associée, *i.e.* telle

que $\mathbb{P}(\sigma = \sigma_0) = f(\sigma_0)$. Elle induit une mesure sur $T(\lambda) \times T(\lambda)$ par :

$$\begin{aligned}\mathbb{P}\left[\sigma \cdot \bar{t} = \overline{u}\right] &= \sum_{\sigma \cdot \bar{t} = \overline{u}} f(\sigma) \\ &= \sum_{\sigma \in \mathfrak{S}_n} f(\sigma) \mathbb{1}_{\left\{\sigma \cdot \bar{t} = \overline{u}\right\}} \\ &= \sum_{\sigma \in \mathfrak{S}_n} f(\sigma) R^{\lambda}(\sigma)_{\overline{u}, \bar{t}} \\ &= \left( \sum_{\sigma \in \mathfrak{S}_n} f(\sigma) \rho_{\lambda}(\sigma) \right)_{\overline{u}, \bar{t}}\end{aligned}$$

On note comme avant :

$$\widehat{f}(\rho_{\lambda}) = \sum_{\sigma \in \mathfrak{S}_n} f(\sigma) \rho_{\lambda}(\sigma)$$

On l'appelle la matrice des $\lambda$-marginales. On définit aussi les fonctions $\lambda$-marginales par :

$$f_{\bar{t}} : T(\lambda) \to \mathbb{C}, \quad \overline{u} \mapsto \sum_{\sigma \cdot \bar{t} = \overline{u}} f(\sigma)$$

Ce sont toutes des distributions de probabilités. On voit que ce sont les vecteurs colonnes de la matrice des $\lambda$-marginales, ce qui justifie son nom.

**Marginales d'ordre $k$** On dit qu'une $\lambda$-marginale (ou une partition $\lambda$) est d'ordre $k$ si $\lambda_1 = n - k$. Ces partitions sont en bijection avec les partitions de $k$, il y en a donc autant, et elles ne sont pas totalement ordonnées. Il y a cependant deux partitions particulières, qui peuvent se comparer à toutes les autres :

- la plus petite : $(n-k, k)$
- la plus grande : $\left(n-k, 1^k\right)$

De plus, la partition $\left(n-k, 1^k\right)$ est plus grande que toutes les partitions d'ordre $j$, pour $j \leqslant k$.

On note $\mathcal{A}_k$ l'ensemble des $k$-arrangements de $\{1, ..., n\}$ (*i.e.* l'ensemble des $k$-uplets sans répétition). On a $|\mathcal{A}_k| = n(n-1)...(n-k+1)$.
On a vu que $T\left(n-k, 1^k\right)$ est en bijection avec $\mathcal{A}_k$. Donc on identifie les fonctions de $M^{\left(n-k, 1^k\right)}$ aux fonctions $\mathcal{A}_k \to \mathbb{C}$.

L'ordre partiel permet en quelque sorte de comparer la quantité d'information contenue dans les marginales. On peut le voir par exemple au sein des marginales d'ordre $k$. Soit $\lambda$ une partition d'ordre $k$. On assimile comme précédemment un $\lambda$-tabloïde $\bar{t} \in T(\lambda)$ à sa partition ordonnée $(A_1, ..., A_r)$, où $\{A_1, ..., A_r\}$ est une partition de $\{1, ..., n\}$ avec $|A_i| = \lambda_i$. On note $x_1, ..., x_{\lambda_2}$ les éléments de $A_2$, $x_{\lambda_2+1}, ..., x_{\lambda_2+\lambda_3}$ ceux de $A_3$, et ainsi de suite.
On pose alors :

$$\begin{aligned}\Pi^{\lambda} : M^{\left(n-k, 1^k\right)} &\to M^{\lambda} \\ f &\mapsto f_{\lambda}\end{aligned}$$

$$\text{avec} \qquad f_{\lambda}(A_1, ..., A_r) = \sum_{\substack{\sigma \in \mathfrak{S}(\{x_1, ..., x_k\}) \\ \sigma \cdot (A_2, ..., A_r) = (A_2, ..., A_r)}} f(\sigma(x_1), ..., \sigma(x_k))$$

## E.2 Fonctions *prolate*

### E.2.1 Généralités

**Définitions et notations**

**Notations :** Pour $f \in \mathbb{C}[\mathfrak{S}_n]$, on note

- $\text{supp}(f) = \{\sigma \in \mathfrak{S}_n |\ f \neq 0\}$ son support
- $\text{supp}\left(\widehat{f}\right) = \left\{\lambda \vdash n |\ \widehat{f}(\lambda) \neq 0\right\}$ le support de sa transformée de Fourier

**Projection dans le domaine "temporel" :** On définit une première opération de projection qui consiste simplement à tronquer une fonction $f \in \mathbb{C}[\mathfrak{S}_n]$ à un sous-ensemble $S \subset \mathfrak{S}_n$ donné :

$$\Pi_S : f \mapsto f\mathbb{1}_S$$

$\Pi_S$ est clairement un projecteur, et on voit facilement qu'il est autoadjoint donc orthogonal.

**Projection dans le domaine "fréquentiel" :** On définit une deuxième opération de projection qui consiste à tronquer une fonction $f \in \mathbb{C}[\mathfrak{S}_n]$ à un sous-ensemble de ses coefficients de Fourier $\Lambda \subset \{\lambda \vdash n\}$ donné. On veut donc :

$$\begin{aligned}\widehat{P_\Lambda f}(\lambda) &= \widehat{f}(\lambda)\mathbb{1}_{\lambda\in\Lambda} \\ &= \mathcal{F}(f * \psi_\Lambda)(\lambda) \quad \text{avec} \quad \widehat{\psi_\Lambda}(\lambda) = I_{d_\lambda}\mathbb{1}_{\lambda\in\Lambda}\end{aligned}$$

On pose donc :

$$P_\Lambda : f \mapsto f * \psi_\Lambda$$

En regardant les transformées de Fourier, on voit facilement que $P_\Lambda$ est un projecteur, et qu'il est autoadjoint donc orthogonal.

**Opérateurs prolate :** Enfin, on définit l'opérateur $A_{S,\Lambda} = P_\Lambda \Pi_S$ que l'on note simplement $A$ s'il n'y a pas d'ambiguïté.
Son opérateur adjoint est alors $A^* = \Pi_S^* P_\Lambda^* = \Pi_S P_\Lambda$.
On a aussi :

$$AA^* = P_\Lambda \Pi_S P_\Lambda \qquad \text{et} \qquad A^*A = \Pi_S P_\Lambda \Pi_S$$

Ces deux derniers opérateurs sont naturellement autoadjoints et positifs. De plus, en notant $|||.|||$ la norme d'opérateur (subordonnée à la norme hermitienne), comme $\Pi_S$ et $P_\Lambda$ sont des projecteurs orthogonaux, on a :

$$|||AA^*||| \leqslant |||P_\Lambda|||||\Pi_S|||||P_\Lambda||| \leqslant 1 \qquad \text{et} \qquad |||A^*A||| = |||\Pi_S|||||P_\Lambda|||||\Pi_S||| \leqslant 1$$

Toutes leurs valeurs propres sont donc dans $[0, 1]$.

**Les *fonctions prolate*** On dispose d'une fonction $f \in \mathbb{C}[\mathfrak{S}_n]$, et on aimerait pouvoir localiser l'information qu'elle contient en temps et en fréquence, au sein d'un produit scalaire. Supposons qu'il existe une fonction $\phi \in \mathbb{C}[\mathfrak{S}_n]$ telle que $\text{supp}(\phi) \subset S$ et $\text{supp}\left(\widehat{\phi}\right) \subset \Lambda$. On a alors :

$$\begin{aligned}\langle f, \phi \rangle &= \sum_{\sigma \in S} \overline{f(\sigma)} \phi(\sigma) \\ &= \frac{1}{n!} \sum_{\lambda \in \Lambda} d_\lambda \langle \widehat{f}(\lambda), \widehat{\phi}(\lambda) \rangle\end{aligned}$$

On voit donc que, connaissant $\phi$, le produit scalaire $\langle f, \phi \rangle$ ne contient que l'information de f sur $S$ et sur $\Lambda$.
Le problème, c'est que $\phi$ doit respecter l'inégalité d'incertitude :

$$|\text{supp}(\phi)| \sum_{\lambda \in \text{supp}\left(\widehat{\phi}\right)} (d_\lambda)^2 \geqslant n!$$

Il n'est donc pas possible de restreindre $\phi$ à $S$ et $\Lambda$ s'ils sont petits. L'idée est alors de chercher $\phi$ restreinte seulement à l'un des deux, mais dont la restriction à l'autre a une norme maximale. On est donc amené à considérer les deux problèmes suivants :

$$\max_{\substack{\|f\|=1 \\ \text{supp}\left(\widehat{f}\right) \subset \Lambda}} \sum_{\sigma \in S} |f(\sigma)|^2 \qquad \text{et} \qquad \max_{\substack{\|f\|=1 \\ \text{supp}(f) \subset S}} \frac{1}{n!} \sum_{\lambda \in \Lambda} d_\lambda \|\widehat{f}(\lambda)\|^2$$

(où $\|.\|$ désigne aussi bien la norme hermitienne sur $\mathbb{C}[\mathfrak{S}_n]$ que la norme de Frobénius sur $\mathcal{M}_d(\mathbb{C})$).
Une fonction qui optimise l'un des deux problèmes est appelée *fonction prolate*. On peut les réécrire avec les opérateurs de projections :

$$\max_{\substack{\|f\|=1 \\ P_\Lambda f = f}} \|\Pi_S f\|^2 \qquad \text{et} \qquad \max_{\substack{\|f\|=1 \\ \Pi_S f = f}} \|P_\Lambda f\|^2$$

Maintenant, le premier se transforme de la façon suivante :

$$\begin{aligned}\max_{\substack{\|f\|=1 \\ P_\Lambda f = f}} \|\Pi_S f\|^2 &= \max_{\substack{\|f\|=1 \\ P_\Lambda f = f}} \|\Pi_S P_\Lambda f\|^2 \\ &= \max_{\|f\|=1} \|\Pi_S P_\Lambda f\|^2 \\ &= \max_{\|f\|=1} \langle A^* f, A^* f \rangle \\ &= \max_{\|f\|=1} \langle A A^* f, f \rangle\end{aligned}$$

(le passage de la première à la deuxième ligne se justifie en disant que si $\phi = \text{argmax}\{\|\Pi_S P_\Lambda f\| \mid \|f\| = 1\}$ alors $\|\Pi_S P_\Lambda \phi\| \geqslant \|\Pi_S P_\Lambda \frac{P_\Lambda \phi}{\|P_\Lambda \phi\|}\|$ *i.e.* $\|P_\Lambda \phi\| \geqslant 1$ donc $P_\Lambda \phi = \phi$)

Ainsi, comme $AA^*$ est autoadjoint positif, $\max_{\substack{\|f\|=1 \\ P_\Lambda f = f}} \|\Pi_S f\|^2$ est égal à sa plus grande valeur propre, et il est atteint pour les vecteurs propres associés.
De même, $\max_{\substack{\|f\|=1 \\ \Pi_S f = f}} \|P_\Lambda f\|^2$ est égal à la plus grande valeur propre de $A^* A$ et est atteint pour les vecteurs propres associés.
Il suffit donc d'appliquer l'algorithme des puissances itérées aux matrices $AA^*$ et $A^* A$ pour résoudre numériquement les deux problèmes initiaux.

**Etude des opérateurs $AA^*$ et $A^*A$** Nous allons maintenant exprimer la matrice de $A^*A$. Soit $\sigma \in \mathfrak{S}_n$. On a :

$$\Pi_S \delta_\sigma = \delta_\sigma \mathbb{1}_{\sigma \in S}$$
$$P_\Lambda \delta_\sigma = \delta_\sigma * \psi_\Lambda = \sum_{\sigma' \in \mathfrak{S}_n} \psi_\Lambda \left(\sigma^{-1}\sigma'\right) \delta_{\sigma'}$$

D'où :

$$A^*A\, \delta_\sigma = \Pi_S P_\Lambda \Pi_S\, \delta_\sigma = \mathbb{1}_{\sigma \in S} \sum_{\sigma' \in S} \psi_\Lambda \left(\sigma^{-1}\sigma'\right) \delta_{\sigma'}$$

On numérote les permutations de $\mathfrak{S}_n$ en commençant avec les permutations de $S$ : $\sigma_1, \sigma_2, ..., \sigma_s$. La matrice de $A^*A$ dans la base canonique est alors :

$$\left(\begin{array}{c|c} M & 0 \\ \hline 0 & 0 \end{array}\right) \qquad \text{avec} \qquad M = \begin{pmatrix} \psi_\Lambda\left(\sigma_1^{-1}\sigma_1\right) & \cdots & \psi_\Lambda\left(\sigma_s^{-1}\sigma_1\right) \\ \vdots & \ddots & \vdots \\ \psi_\Lambda\left(\sigma_1^{-1}\sigma_s\right) & \cdots & \psi_\Lambda\left(\sigma_s^{-1}\sigma_s\right) \end{pmatrix}$$

On peut détailler un peu plus en développant la fonction $\psi_\Lambda$ par la formule d'inversion de Fourier :

$$\psi_\Lambda(\sigma) = \frac{1}{n!} \sum_{\lambda \in \Lambda} d_\lambda \chi_\lambda\left(\sigma^{-1}\right) = \frac{1}{n!} \sum_{\lambda \in \Lambda} d_\lambda \chi_\lambda(\sigma)$$

On peut alors écrire :

$$M = \frac{1}{n!} \sum_{\lambda \in \Lambda} d_\lambda X_\lambda \qquad \text{avec} \qquad X_\lambda = \begin{pmatrix} \chi_\lambda\left(\sigma_1^{-1}\sigma_1\right) & \cdots & \chi_\lambda\left(\sigma_s^{-1}\sigma_1\right) \\ \vdots & \ddots & \vdots \\ \chi_\lambda\left(\sigma_1^{-1}\sigma_s\right) & \cdots & \chi_\lambda\left(\sigma_s^{-1}\sigma_s\right) \end{pmatrix}$$

En ce qui concerne $AA^*$, comme il a le même polynôme caractéristique que $A^*A$, il a les mêmes valeurs propres. De plus, si $\alpha \neq 0$ est une valeur propre, en notant $E_\alpha$ l'espace propre associé pour $A^*A$ et $F_\alpha$ l'espace propre associé pour $AA^*$, on a :

$$\dim E_\alpha = \dim F_\alpha \qquad \text{et} \qquad F_\alpha = A\left(E_\alpha\right)$$

Donc les vecteurs propres de $AA^*$ s'obtiennent facilement à partir de ceux de $A^*A$.

### E.2.2 Propriétés à regarder

**Localisation** Soit $S \subset \mathfrak{S}_n$ et $\Lambda \subset \{\lambda \vdash n\}$. On note :

$$\alpha = \max_{\substack{\|f\|=1 \\ \operatorname{supp}\left(\widehat{f}\right) \subset \Lambda}} \sum_{\sigma \in S} |f(\sigma)|^2 = \max_{\substack{\|f\|=1 \\ \operatorname{supp}(f) \subset S}} \frac{1}{n!} \sum_{\lambda \in \Lambda} d_\lambda \|\widehat{f}(\lambda)\|^2$$

Ainsi que $\phi_1$ une fonction qui atteint le premier max, et $\phi_2$ une fonction qui atteint le deuxième. L'intuition initiale était que $\phi_1$ et $\phi_2$ permettent de localiser au mieux une fonction $f \in \mathbb{C}\left[\mathfrak{S}_n\right]$ à travers ses produits scalaires. On aimerait donc avoir :

$$|\langle \Pi_S f, \Pi_S \phi_1 \rangle| \simeq \max_{\substack{\|g\|=1 \\ \operatorname{supp}(\widehat{g}) \subset \Lambda}} |\langle \Pi_S f, \Pi_S g \rangle|$$
$$|\langle P_\Lambda f, P_\Lambda \phi_2 \rangle| \simeq \max_{\substack{\|g\|=1 \\ \operatorname{supp}(g) \subset S}} |\langle P_\Lambda f, P_\Lambda g \rangle|$$

Mais $\phi_1$ et $\phi_2$ ne dépendent pas de $f$, donc il n'y a pas de raison que ce soit le cas a priori. Par contre, si on suppose que $f$ est une variable aléatoire sur $\mathbb{C}[\mathfrak{S}_n]$, on pourrait avoir un résultat du type :

$$\mathbb{E}[|\langle \Pi_S f, \Pi_S \phi_1 \rangle|] = \max_{\substack{\|g\|=1 \\ \text{supp}(\widehat{g}) \subset \Lambda}} \mathbb{E}[|\langle \Pi_S f, \Pi_S g \rangle|]$$
$$\mathbb{E}[|\langle P_\Lambda f, P_\Lambda \phi_2 \rangle|] = \max_{\substack{\|g\|=1 \\ \text{supp}(g) \subset S}} \mathbb{E}[|\langle P_\Lambda f, P_\Lambda g \rangle|]$$

Sinon, on peut peut-être montrer que la fonction $g$ qui maximise s'écrit comme une combinaison linéaire de tous les vecteurs propres de l'opérateur associé.

**Valeurs propres et vecteurs propres** Dans la même optique, si on suppose que les valeurs propres mesurent la quantité d'information que l'on peut coder, on peut étudier plusieurs chose :

- CNS sur $S$ et $\Lambda$ pour que 1 soit valeur propre (ou valeur propre multiple)
- Etudier la décroissance des valeurs propres en fonction de $S$ et $\Lambda$ (le but étant de savoir dans quels cas une grosse part de l'information est donnée par peu de valeurs propres)

### E.2.3 Applications possibles

**Reconstruction de fonction (problème de Shah) ou approximation** Soit $f \in \mathbb{C}[\mathfrak{S}_n]$ une fonction inconnu que l'on observe partiellement sous plusieurs formes possibles :

- on observe quelques coefficients de Fourier
- on observe quelques valeurs de sa transformée de Radon
- on observe quelques valeurs sur les permutations

On veut la retrouver ou l'approcher en supposant en plus une condition dessus (sinon il y a trop de solutions) :

- elle est sparse
- son gradient est sparse
- elle est à bande limitée

Dans chaque cas, on peut :

- déterminer les types de fonctions pour lesquelles il est possible de les retrouver exactement
- donner une procédure algorithmique d'approximation
- donner des bornes d'approximation (pour la procédure et en général)

**Echantillonnage** On se donne une distance $d$ sur $\mathfrak{S}_n$. On se prend (ou on cherche) un $T \subset \mathfrak{S}_n$ tel que :

$$\begin{aligned} \forall \tau, \tau' \in T, \quad & d(\tau, \tau') > 1 \\ \forall \sigma \in \mathfrak{S}_n, \ \exists \tau \in T, \quad & d(\sigma, \tau) \leqslant 1 \end{aligned}$$

Le but est alors d'échantillonner une fonction $f \in \mathbb{C}[\mathfrak{S}_n]$ en une fonction dont le support est dans $T$.
L'idée est de prendre la fonction $\Pi_T(f * \phi)$, où $\phi$ est la fonction prolate associée à $T$ et un ensemble de fréquences (typiquement les basses fréquences)

**Transformée en ondelettes** Soit $\mathcal{S}$ une partition de $\mathfrak{S}_n$.
Soit $\mathcal{L}$ une partition de $\{\lambda \vdash n\}$.
Pour chaque couple $(S, \Lambda) \in \mathcal{S} \times \mathcal{L}$, on note $\phi^1_{S,\Lambda}, ..., \phi^s_{S,\Lambda}$ (avec $s = |S|$) les vecteurs propres de $A^*A$.
On a alors :

- A $(S, \Lambda)$ fixé, $\left\{\phi^k_{S,\Lambda}\right\}_{1 \leqslant k \leqslant s}$ est une base orthonormée des fonctions de $\mathbb{C}[\mathfrak{S}_n]$ à support dans $S$
- Pour $S \neq S'$, les fonctions sont orthogonales
- La famille totale forme un dictionnaire de $\mathbb{C}[\mathfrak{S}_n]$

Le but est de trouver $\mathcal{S}$ et $\mathcal{L}$ tels que le dictionnaire soit le plus efficace pour la compression de signaux sur $\mathbb{C}[\mathfrak{S}_n]$.

## E.3 Analyse multi-résolution

### E.3.1 Généralités

**Définitions et notations** Soit $n \geqslant 1$.
Pour $k \in [\![1, n-1]\!]$, on note $\mathcal{A}_k = \{(i_1, ..., i_k) \in [\![1, n]\!]^k : i_p \neq i_q \text{ pour } p \neq q\}$.
On note aussi $\mathcal{A}_0 = \{0\}$.
$|\mathcal{A}_k| = \dfrac{n!}{(n-k)!} = n(n-1)...(n-k+1)$
Pour $i = (i_1, ..., i_k)$ et $j = (j_1, ..., j_k)$ des éléments de $\mathcal{A}_k$, on note $\sigma(i) = j$ si $\sigma(i_p) = j_p$ pour tout $p \in [\![1, k]\!]$.
Pour $k \in [\![0, n-1]\!]$, on note $\mathfrak{S}_{n-k}$ le groupe symétrique sur $\{1, ..., n-k\}$, que l'on assimile à $\{\sigma \in \mathfrak{S}_n : \sigma(n-p) = n-p \text{ pour } p \in [\![0, k-1]\!]\}$.
On a ainsi une suite de sous-groupes emboîtés : $\{id\} = \mathfrak{S}_1 \leqslant \mathfrak{S}_2 \leqslant ... \leqslant \mathfrak{S}_n$.

Soit $k \in [\![0, n-1]\!]$.
Pour $i = (i_1, ..., i_k)$ et $j = (j_1, ..., j_k)$ des éléments de $\mathcal{A}_k$, on note :

$$\begin{aligned} A_i &= \{\sigma \in \mathfrak{S}_n : \sigma(n-p+1) = i_p \text{ pour } p \in [\![1, k]\!]\} \\ B_i &= \{\sigma \in \mathfrak{S}_n : \sigma^{-1}(n-p+1) = i_p \text{ pour } p \in [\![1, k]\!]\} \\ C_{i,j} &= \{\sigma \in \mathfrak{S}_n : \sigma(i_p) = j_p \text{ pour } p \in [\![1, k]\!]\} \end{aligned}$$

(On pose $A_0 = B_0 = \mathfrak{S}_n$). $\{A_i : i \in \mathcal{A}_k\}$ et $\{B_i : i \in \mathcal{A}_k\}$ sont des partitions de $\mathfrak{S}_n$ en $|\mathcal{A}_k|$ éléments de cardinal $(n-k)!$.
De plus, $(\{A_i : i \in \mathcal{A}_k\})_{0 \leqslant k \leqslant n-1}$ forme une suite de partitions emboîtées, avec

$$A_i = \bigsqcup_{x \in [\![1,n]\!] \setminus \{i_1,\dots,i_k\}} A_{i_1,\dots,i_k,x}$$

En choisissant pour chaque $i \in \mathcal{A}_k$ un élément $\pi_i \in A_i$, on obtient un système de générateurs pour les classes à gauches de $\mathfrak{S}_{n-k}$ dans $\mathfrak{S}_n$, *i.e.* :

$$\forall i \in \mathcal{A}_k, \qquad A_i = \pi_i \mathfrak{S}_{n-k}$$

De la même façon, les $B_i$ sont les classes à droites de $\mathfrak{S}_{n-k}$ dans $\mathfrak{S}_n$.

**Représentation orthogonale de Young et transformée de Fourier** On prend comme représentation associée la représentation orthogonale de Young, notée $\rho_\lambda$. Elle présente plusieurs avantages :

- Les $\rho_\lambda$ sont des matrices orthogonales
- La règle de branchement de Young s'exprime sans changement de base
- Elle est de plus adaptée à un ordre total simple sur $ST(\lambda)$ dit "last letter sequence"

Soit $k \in [\![0, n-1]\!]$.
Pour $f \in L(\mathfrak{S}_{n-k})$ et $\mu \vdash n$, on note $\widehat{f}(\mu)$ le "coefficient de Fourier" de $f$ en $\mu$ :

$$\widehat{f}(\mu) = \sum_{\sigma' \in \mathfrak{S}_{n-k}} f(\sigma')\rho_\mu(\sigma')$$

Et on définit :

$$\begin{aligned} \mathcal{F}_k : L(\mathfrak{S}_{n-k}) &\to \bigoplus_{\mu \vdash n-k} \mathcal{M}_{d_\mu}(\mathbb{R}) \\ f &\mapsto \bigoplus_{\mu \vdash n-k} \widehat{f}(\mu) \end{aligned}$$

C'est un isomorphisme d'algèbre (isométrique pour la mesure de Plancherel). On a les formules suivantes :

$$\begin{aligned} \widehat{f * g}(\mu) &= \widehat{f}(\mu)\widehat{g}(\mu) \\ \|f\|^2 &= \sum_{\mu \vdash n-k} \frac{d_\mu}{(n-k)!} \|\widehat{f}(\mu)\|^2 \\ \langle f, g \rangle &= \sum_{\mu \vdash n-k} \frac{d_\mu}{(n-k)!} \langle \widehat{f}(\mu), \widehat{g}(\mu) \rangle \\ f(\sigma) &= \sum_{\mu \vdash n-k} \frac{d_\mu}{(n-k)!} Tr\left[\rho_\mu(\sigma^{-1})\widehat{f}(\mu)\right] \end{aligned}$$

**Décomposition suivant une partition** Soit $k \in [\![0, n-1]\!]$.
On choisit un système de représentants pour les classes à gauches $\{\pi_i\}_{i \in \mathcal{A}_k}$.
(Pour $k = 0$, on prend $\pi_0 = id$).
Pour $f \in L(\mathfrak{S}_{n-k})$ et $i \in \mathcal{A}_k$, on définit la fonction :

$$\begin{aligned} f_i : \mathfrak{S}_{n-k} &\to \mathbb{R} \\ \sigma' &\mapsto f(\pi_i \sigma') \end{aligned}$$

C'est en quelque sorte la restriction de $f$ à $A_i$. On définit alors l'application :

$$\begin{aligned} \Phi_k : L(\mathfrak{S}_n) &\to L(\mathfrak{S}_{n-k})^{|\mathcal{A}_k|} \\ f &\mapsto (f_i)_{i \in \mathcal{A}_k} \end{aligned}$$

C'est un isomorphisme d'espace vectoriels et une isométrie :

$$\|f\|^2 = \sum_{i \in \mathcal{A}_k} \|f_i\|^2$$

On définit maintenant l'application qui à $f$ associe toutes les transformées de Fourier des $f_i$ :

$$\begin{aligned} \overline{\mathcal{F}_k} : L(\mathfrak{S}_n) &\to \left( \bigoplus_{\mu \vdash n-k} \mathcal{M}_{d_\mu}(\mathbb{R}) \right)^{|\mathcal{A}_k|} \\ f &\mapsto \left( \bigoplus_{\mu \vdash n-k} \widehat{f_i}(\mu) \right)_{i \in \mathcal{A}_k} \end{aligned}$$

C'est encore un isomorphisme d'espaces vectoriels et une isométrie :

$$\|f\|^2 = \sum_{i \in \mathcal{A}_k} \sum_{\mu \vdash n-k} \frac{d_\mu}{(n-k)!} \|\widehat{f_i}(\mu)\|^2$$

**Analyse temps-fréquence** Soit $f \in L(\mathfrak{S}_n)$. Chaque $k \in [\![0, n-1]\!]$ donne une partition de "l'espace des temps" $\mathfrak{S}_n$, et les partitions sont emboîtées. Donc plus on augmente $k$, plus on raffine la décomposition de $f$.
Pour $k \in [\![0, n-1]\!]$ et $i \in \mathcal{A}_k$ donnés, $f_i$ contient "l'information temporelle locale" de $f$ sur $A_i$.
Chaque coefficient de Fourier $\widehat{f_i}(\mu)$ contient alors "l'information fréquentielle d'ordre $\mu$" de $f_i$.
Il contient donc une information localisée en temps et en fréquence.
Grâce aux propriétés d'isométrie, on peut écrire :

$$\begin{aligned} \|f\|^2 &= \sum_{\lambda \vdash n} \frac{d_\lambda}{(n)!} \|\widehat{f}(\lambda)\|^2 && \text{(à l'ordre 0)} \\ &\vdots \\ &= \sum_{i \in \mathcal{A}_k} \sum_{\mu \vdash n-k} \frac{d_\mu}{(n-k)!} \|\widehat{f_i}(\mu)\|^2 && \text{(à l'ordre } k\text{)} \\ &\vdots \\ &= \sum_{\sigma \in \mathfrak{S}_n} f(\sigma)^2 && \text{(à l'ordre } n-1\text{)} \end{aligned}$$

A l'ordre $k$, $\|f\|^2$ se décompose en une somme de $|\mathcal{A}_k| \cdot p(n-k)$ termes.

On choisit pour chaque $k \in [\![0, n-1]\!]$ un système de représentants des classes à gauches $\{\pi_i\}_{i\in\mathcal{A}_k}$.
Pour $k \in [\![0, n-1]\!]$, $i \in \mathcal{A}_k$ et $\mu \vdash n-k$, on pose :

$$\begin{aligned}\phi^k_{i,\mu} : \mathfrak{S}_n &\to \mathcal{M}_{d_\mu}(\mathbb{R}) \\ \sigma &\mapsto \mathbb{1}_{A_i}(\sigma)\rho_\mu(\pi^{-1}\sigma)\end{aligned}$$

On a en particulier :

- pour $k = 0$ : $\phi^0_{0,\lambda} = \rho_\lambda$
- pour $k = n-1$ : $\phi^{n-1}_{\sigma,(1)} = \delta_\sigma$

La décomposition précédente de $f$ à l'ordre $k$ correspond donc à la décomposition de $f$ sur la famille $\left(\phi^k_{i,\mu}\right)_{i\in\mathcal{A}_k,\mu\vdash n-k}$, qui est une base de l'espace $\left(\bigoplus_{\mu\vdash n-k}\mathcal{M}_{d_\mu}(\mathbb{R})\right)^{|\mathcal{A}_k|}$, isomorphe à $L(\mathfrak{S}_n)$.
On dispose donc d'un dictionnaire $\{\left(\phi^k_{i,\mu}\right)_{i\in\mathcal{A}_k,\mu\vdash n-k} : k \in [\![0, n-1]\!]\}$, permettant de localiser en temps et en fréquence, par des "produits scalaires", l'information de $f$.

L'idée est donc de construire une procédure d'approximation consistant à calculer les $\sum_{k=0}^{n-1}|\mathcal{A}_k|\cdot p(n-k)$ termes $\frac{d_\mu}{(n-k)!}\|\widehat{f}_i(\mu)\|^2$ et à n'en garder que certains.

## E.3.2 Quelques propriétés

**Décomposition temporelle et transformée de Fourier** Soit $k \in [\![0, n-1]\!]$. Soit $\lambda \vdash n$ et $\mu \vdash n-k$.
On dit que $\lambda$ domine $\mu$ si son diagramme contient celui de $\mu$. On écrit $\lambda \succcurlyeq \mu$.
On choisit un système de représentants des classes à gauches $\{\pi_i\}_{i\in\mathcal{A}_k}$.
On a alors, pour tout $f \in \mathrm{L}(\mathfrak{S}_n)$,

$$\widehat{f}(\rho_\lambda) = \sum_{i\in\mathcal{A}_k}\rho_\lambda(\pi_i)\bigoplus_{\substack{\mu\vdash n-k\\ \mu\preccurlyeq\lambda}}\widehat{f}_i(\rho_\mu)$$

**Calcul des normes** Soit $k \in [\![0, n-1]\!]$. Pour $\mu \vdash n-k$, on définit la matrice $X^\mu \in \mathcal{M}_{(n-k)!}(\mathbb{R})$ par $X^\mu_{\sigma,\tau} = \chi_\mu(\sigma\tau^{-1})$.
On a alors, pour tout $f \in \mathrm{L}(\mathfrak{S}_{n-k})$,

$$\begin{aligned}\|\widehat{f}(\rho_\mu)\|^2 &= \sum_{\sigma,\tau\in\mathfrak{S}_{n-k}} f(\sigma)f(\tau)\chi_\mu(\sigma\tau^{-1}) \\ &= f^T X^\mu f\end{aligned}$$

## E.3.3 Algorithme de Coifman-Wickerhauser

**Principe général** On se place dans l'espace $\mathbb{C}^N$.
On suppose qu'on peut le décomposer de manière récursive en une suite de décomposition orthogonales emboîtées. On représente cette suite de décompositions par un arbre, où chaque nœud

représente un sous-espace.
Pour chaque sous-espace on dispose d'une base orthonormée. On en fait l'union pour obtenir ainsi un dictionnaire de fonctions normées $\mathcal{D} = \{\phi_p\}_{p\in\Gamma}$.
Soit alors $f \in \mathbb{C}^N$. On veut trouver la base qui soit optimale en un certain sens pour représenter $f$ (typiquement pour obtenir la représentation la plus *sparse* possible).

On choisit une fonction de coût $C : \mathbb{R}^+ \to \mathbb{R}^+$, et on définit le coût de $f$ dans la base $\mathcal{B} = \{\phi_p\}_{p\in\Gamma_\mathcal{B}}$ par

$$\mathcal{C}(f, \mathcal{B}) = \sum_{p\in\Gamma_\mathcal{B}} C\left(|\langle f, \phi_p\rangle|\right)$$

on veut alors résoudre le problème d'optimisation :

$$\min_{\mathcal{B}\subset\mathcal{D}} \mathcal{C}(f, \mathcal{B})$$

Si on veut que la solution de ce problème donne une représentation *sparse*, il vaut donc mieux choisir $C$ telle que $C(0) = 0$.

L'algorithme :

1. Initialisation : Calcul de tous les $\mathcal{C}(f, \mathcal{B})$ pour les bases $\mathcal{B}$ associées aux nœuds

2. Actualisation : On démarre avec la base composée de la réunion des bases associées aux feuilles de l'arbre. Pour chaque nœud supérieur, on compare le coût de sa base associée avec la somme des coûts des bases associées aux nœuds fils. S'il est inférieur, on remplace par la base, sinon on garde.

Cet algorithme résout le problème d'optimisation précédent.

**Dans notre cas** L'espace global est $L(\mathfrak{S}_n)$.
Pour $k \in [\![0, n-1]\!]$ et $i \in \mathcal{A}_k$, on note $V_i = \{f \in L(\mathfrak{S}_n) : \mathrm{supp}(f) \subset A_i\}$.
Les $(V_i)_{i,k}$ forment une suite de décompositions orthogonales emboîtées.
Chaque $V_i$ est muni de sa "base de Fourier" $\mathcal{B}_i = \left(\phi^k_{i,\mu}\right)_{\mu\vdash n-k}$.
Pour une fonction $f$ donnée, on veut trouver la base composée d'éléments de $\{\phi^k_{i,\mu} : i \in \mathcal{A}_k,\ \mu \vdash n-k,\ k \in [\![0, n-1]\!]\}$ dans laquelle $f$ admet la décomposition (matricielle) la plus *sparse*.

Un choix possible pour la fonction de coût est de prendre l'entropie de la décomposition dans une base : pour $k \in [\![0, n-1]\!]$ et $i \in \mathcal{A}_k$,

$$\mathcal{C}(f, \mathcal{B}_i) = -\sum_{\mu\vdash n-k} \frac{d_\mu}{(n-k)!}\|\widehat{f}_i(\mu)\|^2 \log\left(\frac{d_\mu}{(n-k)!}\|\widehat{f}_i(\mu)\|^2\right)$$

A la fin de la procédure, la base obtenue est de la forme $\bigcup_{k\in K}\bigcup_{i\in I_k}\{\phi^k_{i,\mu}\}_{\mu\vdash n-k}$, où $K \subset [\![0, n-1]\!]$ et $I_k \subset \mathcal{A}_k$.
Par transformée de Fourier inverse, $f$ se décompose :

$$f(\sigma) = \sum_{k\in K}\sum_{i\in I_k}\sum_{\mu\vdash n-k} \frac{d_\mu}{(n-k)!} Tr\left[\rho_\mu(\sigma^{-1})\widehat{f}_i(\rho_\mu)\right]$$

# E.4 Localisation de l'information

## E.4.1 Introduction

Notre objectif est de construire une analyse multirésolution et une base d'ondelettes qui permettent de "localiser l'information d'une fonction $f \in L(\mathfrak{S}_n)$ sur des sous-groupes d'objets". Avec le recul, je pense qu'on peut donner deux sens à cette expression :

1. Décomposer $f$ sur des composantes $\psi$ constantes par morceaux sur les ensembles $\mathfrak{S}_n(\pi)$ de permutations qui étendent un ranking, localisées sur les rankings d'un sous-ensemble d'objets (dans l'idée d'une décomposition de Haar)
2. Décomposer $f$ sur des composantes $\psi$ qui localisent l'information qu'on va récupérer de $f$ en prenant les produits scalaires avec les indicatrices $\mathbb{1}_{\mathfrak{S}_n(\pi)}$, c'est-à-dire l'information contenue dans les marginales de $f$.

Ces deux points de vue mènent à des constructions similaires mais différentes, sauf pour la localisation en "échelle". Jusqu'à maintenant, on n'a considéré que la première construction (mais avec un point de vue un peu flou entre les deux). La deuxième est nouvelle. On va voir que la différence se situe en fait simplement dans la façon d'injecter $L(\mathfrak{S}_A)$ dans $L(\mathfrak{S}_n)$. La première construction va correspondre à l'opération élémentaire (définie sur les chaînes)

$$\phi_b : \pi \mapsto \frac{1}{|\pi|+1} \sum_{i=1}^{|\pi|+1} \pi \lhd_i b,$$

où $\pi \lhd_i b$ est le mot obtenu en insérant $b$ à la $i^{\text{ème}}$ place dans $\pi$, et la deuxième à

$$\tilde{\phi}_b : \pi \mapsto \frac{1}{2}(\pi \lhd_1 b + \pi \lhd_{|\pi|+1} b) = \frac{1}{2}(b\pi + \pi b).$$

Je rappelle quelques notations pour la suite : on note $\Gamma_A^k$ l'ensemble des mots injectifs de taille $k$ sur $A$ et $\Gamma_A = \bigcup_{k=1}^{|A|} \Gamma_A^k$ (on note $\Gamma_n^k$ et $\Gamma_n$ au lieu de $\Gamma_{[\![n]\!]}^k$ et $\Gamma_{[\![n]\!]}$). On définit aussi les espaces de chaînes (les fonctions sur les mots) $C_k(\Gamma_A) = \{x : \Gamma_A^k \to \mathbb{R}\}$, et $C(\Gamma_A) = \bigoplus_{k=1}^{|A|} C_k(\Gamma_A)$. Le Dirac en $\pi$ est encore noté $\pi$. L'opérateur de suppression est défini sur les Diracs par $\varrho_a : \pi \mapsto \pi \setminus \{a\}$.

## E.4.2 Comparaison des décompositions multirésolution

**Localisation en échelle.**

1. Dans le premier point de vue, on veut décomposer sur des fonctions constantes sur les ensembles $\mathfrak{S}_n(\pi)$. On définit donc naturellement l'espace de résolution $k$ par

$$V^k = \text{span}\,\{\mathbb{1}_{\mathfrak{S}_n(a_1 \prec \ldots \prec a_k)} \mid a_1, ..., a_k \in [\![n]\!],\ ai \neq a_j\}.$$

2. Dans le deuxième point de vue, on veut isoler l'information contenu dans les marginales. On commence donc par définir l'opérateur de projection sur les marginales d'ordre $k$ par

$$M_k : L(\mathfrak{S}_n) \to \bigoplus_{A \in \mathcal{P}_k([\![n]\!])} L(\mathfrak{S}_A)$$
$$f \mapsto (f_A)_{A \in \mathcal{P}_k([\![n]\!])},$$

où $f_A(\pi) = \langle f, \mathbb{1}_{\mathfrak{S}_n(\pi)} \rangle$, puis on définit l'espace de résolution $k$ par

$$V^k = (\ker M_k)^\perp.$$

Il se trouve que ces deux définitions coïncident (ce qui rend la différence entre les deux points de vue moins évidente). Ensuite on fait la construction classique $V^{k+1} = V^k \overset{\perp}{\oplus} W^{k+1}$. Dans les deux cas, l'espace $W^k$ représente l'information gagnée à l'échelle $k$.

**Localisation en objets.**

1. Au sein de l'espace $W^k$, on localise les fonctions qui ne font intervenir les indicatrices de $\mathfrak{S}_n(\pi)$ que pour les rankings $\pi$ relatifs à un sous-ensemble d'objets $A$. On définit donc $W_A^k = W^k \cap V_A^k$, où $V_A^k = \text{span}\{\mathbb{1}_{\mathfrak{S}_n(\pi)} \mid \pi \in \mathfrak{S}_A\}$. On montre ensuite que $W_A^k$ est simplement caractérisé par

$$W_A^k = \{f \in V_A^k \mid f \perp V_B^k \text{ pour tout } B \subsetneq A\},$$

grâce à la grosse formule combinatoire de transfert que j'ai démontrée. C'est cette propriété qui montre que pour tout sous-ensemble d'objets $A$, $W_A^k$ est isomorphe à l'espace $H^k$ défini sur les chaînes par

$$H^k = \{x \in C_k(\Gamma_k) \mid \varrho_a x = 0 \text{ for all } a \in [\![k]\!]\},$$

à travers l'application

$$\begin{aligned} \phi : C(\Gamma_n) &\to L(\mathfrak{S}_n) \\ \pi &\mapsto \frac{|\pi|!}{n!}\mathbb{1}_{\mathfrak{S}_n(\pi)}. \end{aligned}$$

Si $\pi \in \mathfrak{S}_A$ et $[\![n]\!] \setminus A = \{b_1, ..., b_l\}$, on a $\phi(\pi) = \phi_{b_l} \circ ... \circ \phi_{b_1}(\pi)$, où les $\phi_{b_i}$ commutent. On utilise ensuite le résultat de Reiner (dans l'article de memoirs of AMS) qui dit que $\dim H^k = d_k$, ce qui démontre la décomposition $L(\mathfrak{S}_n) = V^0 \oplus \bigoplus_{k=2}^n \bigoplus_{|A|=k} W_A^k$.

2. Au sein de l'espace $W^k$, on localise les fonctions qui ont toutes leurs marginales nulles sur les sous-ensembles d'objets $B$ qui ne contiennent pas $A$. On définit donc l'espace

$$\tilde{W}_A^k = \{f \in W^k \mid f \perp V_B^k \text{ pour tout } B \not\supseteq A\}.$$

Ce que j'ai remarqué, c'est qu'on a aussi une relation du type $\tilde{W}_A^k = W^k \cap \tilde{V}_A^k$, si on définit bien l'espace $\tilde{V}_A^k$. Pour ça, notons $\mathfrak{S}_n[\pi]$ l'ensemble de toutes les permutations $\sigma \in \mathfrak{S}_n$ qui "contiennent" $\pi$. Rigoureusement, on le définit en termes de permutations par

$$\begin{aligned} \mathfrak{S}_n[\pi] = \{\sigma \in \mathfrak{S}_n \mid \exists i_0 \in \{1, ..., n-k+1\}, \\ \sigma^{-1}(i_0) = \pi^{-1}(i_0), ..., \sigma^{-1}(i_0+k-1) = \pi^{-1}(i_0+k-1)\}, \end{aligned}$$

ou en termes de mots injectifs par

$$\mathfrak{S}_n[\pi] = \{\sigma \in \Gamma_n^n \mid \exists \omega_1, \omega_2 \in \Gamma_n \cup \{\overline{0}\}\},\ \sigma = \omega_1 \pi \omega_2\},$$

où $\overline{0}$ désigne le mot vide. Par exemple,

$$\mathfrak{S}_5[123] = \{45123, 54123, 41235, 51234, 12345, 12354\}.$$

On définit alors $\tilde{V}_A^k = \text{span}\{\mathbb{1}_{\mathfrak{S}_n[\pi]} \mid \pi \in \mathfrak{S}_A\}$, et on montre que $\tilde{W}_A^k = W^k \cap \tilde{V}_A^k$ et même que

$$\tilde{W}_A^k = \{f \in \tilde{V}_A^k \mid f \perp V_B^k \text{ pour tout } B \subsetneq A\}$$

(la démonstration repose sur le fait que les opérations $\pi \mapsto b\pi$ et $\pi \mapsto \pi b$ commutent avec l'opération de suppression $\varrho_a$ pour $a \neq b$). On obtient alors que $\tilde{W}_A^k$ est isomorphe à $H^k$ à travers l'application

$$\begin{aligned} \tilde{\phi} : C(\Gamma_n) &\to L(\mathfrak{S}_n) \\ \pi &\mapsto \frac{1}{(n-|\pi|+1)!} \mathbb{1}_{\mathfrak{S}_n[\pi]} \end{aligned}$$

(il est facile de voir que $|\mathfrak{S}_n[\pi]| = (n-k+1)!$). La démonstration de cette deuxième décomposition multirésolution repose alors exactement sur les mêmes résultats mathématiques (à savoir celui de Reiner), et on a pareil $L(\mathfrak{S}_n) = V^0 \oplus \bigoplus_{k=2}^n \bigoplus_{|A|=k} \tilde{W}_A^k$.

**Synthèse : comparaison des structures multirésolution.** Les deux décompositions ont la même structure "verticale" de décomposition en échelle

$$L(\mathfrak{S}_n) = V^0 \oplus \bigoplus_{k=2}^{n} W^k,$$

mais ont des structures "horizontales" de décomposition en objets différentes, la première étant localisée par rapport aux indicatrices des ensembles $\mathfrak{S}_n(\pi)$

$$W^k = \bigoplus_{|A|=k} W^k \cap V_A^k,$$

alors que la deuxième l'est par rapport aux indicatrices des ensembles $\mathfrak{S}_n[\pi]$

$$W^k = \bigoplus_{|A|=k} W^k \cap \tilde{V}_A^k.$$

Dans les deux cas, la décomposition est orthogonale en échelle mais pas en objets.

### E.4.3 Comparaison des ondelettes

La différence entre les deux constructions réside juste dans la façon d'injecter les chaînes sur les mots injectifs dans les fonctions sur les permutations. Tous les résultats qui concernent les chaînes sont donc valables dans les deux cas. C'est le cas de l'algorithme de construction de nos ondelettes. Plus rigoureusement, notons $x_\tau$ la chaîne obtenue par l'algorithme pour la permutation $\tau$, qui est donc un dérangement sur le complémentaire de l'ensemble de ses points fixes, que je note $A$. L'ondelette correspondante est alors définie par :

1. dans la première construction,

$$\psi_\tau = \sum_{\pi \in \mathfrak{S}_A} x_\tau(\pi) \mathbb{1}_{\mathfrak{S}_n(\pi)},$$

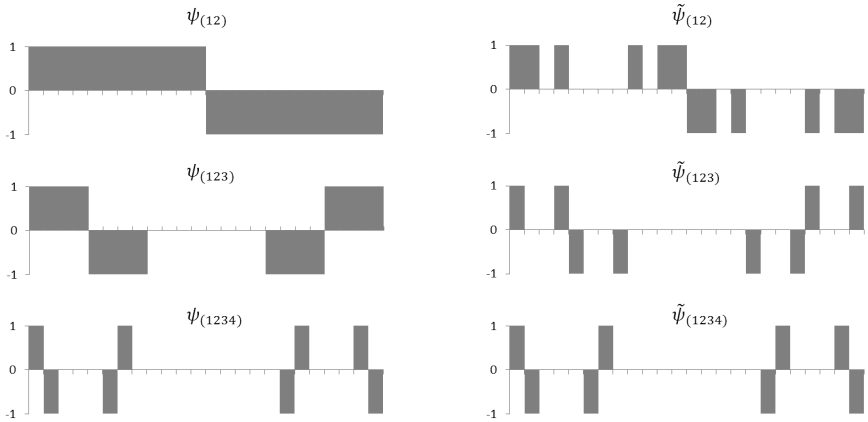
2. dans la deuxième construction

$$\tilde{\psi}_\tau = \sum_{\pi \in \mathfrak{S}_A} x_\tau(\pi) \mathbb{1}_{\mathfrak{S}_n[\pi]}.$$

**Forme des ondelettes.** Si $\tau \in \mathfrak{S}_n$ a $k$ points non fixes et $r$ cycles, $|\operatorname{supp} x_\tau| = 2^{k-r}$. On a donc

$$|\operatorname{supp}(\psi_\tau)| = 2^{k-r}\frac{n!}{k!} \qquad \text{et} \qquad \left|\operatorname{supp}\left(\tilde{\psi}_\tau\right)\right| = 2^{k-r}(n-k+1)!.$$

Dans la première construction, si $\gamma = (a_1 \dots a_k)$ et $b \notin \{a_1, ..., a_k\}$, alors pour tout $j \in \{1, ..., k\}$, $\operatorname{supp}(\psi_{\gamma\cdot(b\,a_j)}) \subset \operatorname{supp}(\psi_\gamma)$. Donc au sein d'un chemin de cycles $(a_1\,a_2), ..., (a_1 \dots a_k)$, les ondelettes $\psi$ associées ont une forme d'ondelettes de Haar combinatoires. Ce n'est pas le cas des ondelettes $\tilde{\psi}$. La figure suivante montre la forme des ondelettes associées aux cycles (12), (123) et (1234) dans les deux constructions.




**Calcul de la transformée en ondelettes.** J'appelle transformée en ondelettes d'une fonction $f \in L(\mathfrak{S}_n)$ l'ensemble des produits scalaires avec les ondelettes (c'est apparemment la terminologie consacrée, même dans le cas d'ondelettes non orthogonales). Dans les deux cas, le calcul de la transformée se fait d'abord sur les ondelettes indexées par des cycles, puis sur celles indexées par des permutations à au moins deux cycles. Ce qui change, ce sont les ondelettes qui vont intervenir. Soit $\sigma \in \mathfrak{S}_n$ et $\tau \in \mathfrak{S}_n$ avec $\operatorname{supp}(\tau) = A$. Le produit scalaire du Dirac $\delta_\sigma$ avec l'ondelette associée à $\tau$ est égal à

1. $\langle \delta_\sigma, \psi_\tau \rangle = \sum_{\pi \in \mathfrak{S}_A} x_\tau(\pi) \mathbb{1}_{\mathfrak{S}_n(\pi)}(\sigma)$ dans la première construction,
2. $\langle \delta_\sigma, \psi_\tau \rangle = \sum_{\pi \in \mathfrak{S}_A} x_\tau(\pi) \mathbb{1}_{\mathfrak{S}_n[\pi]}(\sigma)$ dans la deuxième.

Or, pour tout $A$, il existe un unique $\pi \in \mathfrak{S}_A$ tel que $\mathbb{1}_{\mathfrak{S}_n(\pi)}(\sigma) \neq 0$ (c'est $\sigma_{|A}$), alors qu'il n'existe un $\pi \in \mathfrak{S}_A$ tel que $\mathbb{1}_{\mathfrak{S}_n[\pi]}(\sigma) \neq 0$ que si tous les éléments de $A$ sont collés dans l'écriture de $\sigma$, auquel cas il est unique. Pour $i, j \in \{1, ..., n\}$ avec $i < j$, notons $\sigma_{[\![i,j]\!]} = \sigma^{-1}(i)...\sigma^{-1}(j)$ le sous-mot contigu de $\sigma$ entre les rangs $i$ et $j$, et $\sigma^{-1}([\![i, j]\!]) = \{\sigma^{-1}(i), ..., \sigma^{-1}(j)\}$. Le calcul de la transformée en ondelettes de $\delta_\sigma$ se fait alors à chaque étape

1. en énumérant tous les $\sigma_{|A}$ pour $A \subset [\![n]\!]$, $2 \leq |A| \leq n$, dans la première construction,
2. en énumérant tous les $\sigma_{[\![i,j]\!]}$ pour $1 \leq i < j \leq n$, dans la deuxième construction.

Ce qui mène aux complexités suivantes, en notant

1. $N_A(\sigma)$ le nombre de cycles $\gamma \in \operatorname{Cycle}(A)$ tels que $\psi_\gamma(\sigma_{|A}) \neq 0$

2. $N_{i,j}(\sigma)$ le nombre de cycles $\gamma \in \mathrm{Cycle}(\sigma^{-1}([\![i,j]\!]))$ tels que $\psi_\gamma(\sigma_{[\![i,j]\!]}) \neq 0$.

1. **Première construction**

   - Pour les ondelettes indexées par des cycles

$$\sum_{k=2}^{n} \sum_{|A|=k} N_A(\sigma) \leq \sum_{k=2}^{n} \binom{n}{k} 2^{k-2} = O(3^n).$$

   - Pour les ondelettes indexées par des produits de cycles

$$\sum_{k=4}^{n} \sum_{|A|=k} \sum_{r=2}^{\lfloor k/2 \rfloor} \sum_{\mathbf{k} \in \Sigma_r(k)} \mathbb{I}\{\mathcal{I}_{\mathbf{k}}(\sigma_{|A}) \in \mathrm{Stand}_r(A)\} \prod_{i=1}^{r} N_{A_i}(\sigma)$$
$$\leq \sum_{k=4}^{n} \sum_{|A|=k} \sum_{r=2}^{\lfloor k/2 \rfloor} \binom{k-r-1}{r-1} 2^{k-2r} \leq O\left(\left(\frac{7}{2}\right)^n\right).$$

2. **Deuxième construction**

   - Pour les ondelettes indexées par des cycles

$$\sum_{1 \leq i < j \leq n} N_{i,j}(\sigma) \leq \sum_{1 \leq i < j \leq n} 2^{j-i-1} = O(2^n).$$

   - Pour les ondelettes indexées par des produits de cycles

$$\sum_{k=4}^{n} \sum_{i=1}^{n-k+1} \sum_{r=2}^{\lfloor k/2 \rfloor} \sum_{\mathbf{k} \in \Sigma_r(k)} \mathbb{I}\{\mathcal{I}_{\mathbf{k}}(\sigma_{[\![i,j]\!]}) \in \mathrm{Stand}_r(\sigma^{-1}([\![i,j]\!]))\} \prod_{i=1}^{r} N_{A_i}(\sigma)$$
$$\leq \sum_{k=4}^{n} \sum_{i=1}^{n-k+1} \sum_{r=2}^{\lfloor k/2 \rfloor} \binom{k-r-1}{r-1} 2^{k-2r} \leq O\left(\left(\frac{5}{2}\right)^n\right).$$

### E.4.4 Localisation de l'information

La principale différence entre les deux constructions est bien sûr la façon de localiser l'information. Regardons dans le cas où $n = 3$. Le tableau suivant donne les deux bases d'ondelettes (non normalisées),

| | $\psi_{id}$ | $\psi_{(12)}$ | $\psi_{(13)}$ | $\psi_{(23)}$ | $\psi_{(123)}$ | $\psi_{(132)}$ | $\tilde{\psi}_{id}$ | $\tilde{\psi}_{(12)}$ | $\tilde{\psi}_{(13)}$ | $\tilde{\psi}_{(23)}$ | $\tilde{\psi}_{(123)}$ | $\tilde{\psi}_{(132)}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 123 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 |
| 132 | 1 | 1 | 1 | -1 | -1 | 1 | 1 | 0 | 1 | -1 | -1 | 1 |
| 213 | 1 | -1 | 1 | 1 | 0 | -1 | 1 | -1 | 1 | 0 | 0 | -1 |
| 231 | 1 | -1 | -1 | 1 | -1 | 1 | 1 | 0 | -1 | 1 | -1 | 1 |
| 312 | 1 | 1 | -1 | -1 | 0 | -1 | 1 | 1 | -1 | 0 | 0 | -1 |
| 321 | 1 | -1 | -1 | -1 | 1 | 0 | 1 | -1 | 0 | -1 | 1 | 0 |

et celui-là donne les coefficients de décomposition pour une probabilité $p$ sur $\mathfrak{S}_3$

| | $\Psi$ | $\tilde{\Psi}$ |
|---|---|---|
| $id$ | $\frac{1}{6}$ | $\frac{1}{6}$ |
| $(12)$ | $\frac{1}{4}\left(2\mathbb{P}\left[1 \prec 2\right] - \mathbb{P}\left[1 \prec 3\right] + \mathbb{P}\left[2 \prec 3\right] - 1\right)$ | $\frac{1}{2}\left(\mathbb{P}\left[1 \prec 2\right] - \frac{1}{2}\right)$ |
| $(13)$ | $\frac{1}{4}\left(2\mathbb{P}\left[1 \prec 3\right] - \mathbb{P}\left[1 \prec 2\right] - \mathbb{P}\left[2 \prec 3\right]\right)$ | $\frac{1}{2}\left(\mathbb{P}\left[1 \prec 3\right] - \frac{1}{2}\right)$ |
| $(23)$ | $\frac{1}{4}\left(2\mathbb{P}\left[2 \prec 3\right] + \mathbb{P}\left[1 \prec 2\right] - \mathbb{P}\left[1 \prec 3\right] - 1\right)$ | $\frac{1}{2}\left(\mathbb{P}\left[2 \prec 3\right] - \frac{1}{2}\right)$ |
| $(123)$ | $\frac{1}{2}\left(\mathbb{P}\left[1 \prec 2 \prec 3 \text{ ou } 3 \prec 2 \prec 1\right] - \frac{1}{3}\right)$ | $\frac{1}{2}\left(\mathbb{P}\left[1 \prec 2 \prec 3 \text{ ou } 3 \prec 2 \prec 1\right] - \frac{1}{3}\right)$ |
| $(132)$ | $\frac{1}{2}\left(\frac{1}{3} - \mathbb{P}\left[2 \prec 1 \prec 3 \text{ ou } 3 \prec 1 \prec 2\right]\right)$ | $\frac{1}{2}\left(\frac{1}{3} - \mathbb{P}\left[2 \prec 1 \prec 3 \text{ ou } 3 \prec 1 \prec 2\right]\right)$ |

Les coefficients d'ordre 3 sont les mêmes puisque dans les deux cas ya pas d'injection. Par contre, on voit que les coefficients d'ordre 2 sont différents : ils font intervenir les trois probabilités $\mathbb{P}\left[1 \prec 2\right]$, $\mathbb{P}\left[1 \prec 3\right]$ et $\mathbb{P}\left[2 \prec 3\right]$ dans la première construction alors qu'ils sont complètement localisés dans la deuxième. Mais dans chaque construction, l'ensemble des coefficients d'ordre 2 localise toute l'information d'ordre 2 (les deux structures verticales étant les mêmes). En l'occurrence, notons

$$\begin{aligned} p &= c_{id}\psi_{id} + c_{(12)}\psi_{(12)} + c_{(13)}\psi_{(13)} + c_{(23)}\psi_{(23)} + c_{(123)}\psi_{(123)} + c_{(132)}\psi_{(132)} \\ &= \tilde{c}_{id}\tilde{\psi}_{id} + \tilde{c}_{(12)}\tilde{\psi}_{(12)} + \tilde{c}_{(13)}\tilde{\psi}_{(13)} + \tilde{c}_{(23)}\tilde{\psi}_{(23)} + \tilde{c}_{(123)}\tilde{\psi}_{(123)} + \tilde{c}_{(132)}\tilde{\psi}_{(132)}. \end{aligned}$$

Supposons qu'on connaisse parfaitement $p$ jusqu'à l'ordre 2. On peut alors l'estimer par

1. $p_1 = c_{id}\psi_{id} + c_{(12)}\psi_{(12)} + c_{(13)}\psi_{(13)} + c_{(23)}\psi_{(23)}$ dans le premier cas,
2. $p_2 = \tilde{c}_{id}\tilde{\psi}_{id} + \tilde{c}_{(12)}\tilde{\psi}_{(12)} + \tilde{c}_{(13)}\tilde{\psi}_{(13)} + \tilde{c}_{(23)}\tilde{\psi}_{(23)}$ dans le deuxième.

Si maintenant on veut estimer $\mathbb{P}\left[2 \prec 1 \prec 3\right]$ par exemple, on obtient

$$\begin{aligned} \langle p_1, \delta_{213} \rangle &= c_{id} - c_{(12)} + c_{(13)} + c_{(23)} \\ &= \frac{1}{6} + \frac{1}{2}\left(-\mathbb{P}\left[1 \prec 2\right] + \mathbb{P}\left[1 \prec 3\right]\right) \end{aligned}$$

dans le premier cas et

$$\begin{aligned} \langle p_2, \delta_{213} \rangle &= \tilde{c}_{id} - \tilde{c}_{(12)} + \tilde{c}_{(13)} \\ &= \frac{1}{6} + \frac{1}{2}\left(-\mathbb{P}\left[1 \prec 2\right] + \mathbb{P}\left[1 \prec 3\right]\right) \end{aligned}$$

dans le deuxième. On obtient donc la même chose, mais en sommant moins de coefficients. Plus généralement, pour une probabilité $p$ sur $\mathfrak{S}_n$, en notant toujours $c_\tau$ et $\tilde{c}_\tau$ les coefficients dans la base associée, on a, après calcul, $c_{id} = \tilde{c}_{id} = 1/(n!)$ et pour $1 \leq i < j \leq n$,

$$\begin{aligned} c_{(i\,j)} &= \frac{1}{n+1}\left((n-1)\mathbb{P}[i \prec j] + \sum_{s<i}\mathbb{P}[s \prec i] - \sum_{\substack{s>i \\ s\neq j}}\mathbb{P}[i \prec s] - \sum_{\substack{r<j \\ r\neq i}}\mathbb{P}[r \prec j] + \sum_{r>j}\mathbb{P}[j \prec r]\right) \\ \tilde{c}_{(i\,j)} &= \frac{1}{(n-1)!}\left(\mathbb{P}[i \prec j] - \frac{1}{2}\right), \end{aligned}$$

ce qui donne encore

$$\left\langle p_1, \mathbb{1}_{\mathfrak{S}_n(2\prec 1\prec 3)} \right\rangle = \left\langle p_2, \mathbb{1}_{\mathfrak{S}_n(2\prec 1\prec 3)} \right\rangle = \frac{1}{6} + \frac{1}{2}\left(-\mathbb{P}\left[1 \prec 2\right] + \mathbb{P}\left[1 \prec 3\right]\right).$$

Cela montre que l'information d'ordre 2 utile pour estimer $2 \prec 1 \prec 3$ est uniquement contenue dans $\mathbb{P}[1 \prec 2]$ et $\mathbb{P}[1 \prec 3]$. Or, dans la première construction, on la récupère en faisant la somme de tous les coefficients $c_{(i\,j)}$ tels que $\left\langle \psi_{(i\,j)}, \mathbb{1}_{\mathfrak{S}_n(2\prec 1\prec 3))} \right\rangle \neq 0$, à savoir tous les $c_{(i\,j)}$ avec $\{i,j\} \cap \{1,2,3\} \neq \emptyset$, soit $3n-8$ coefficients, alors que dans la deuxième construction, on la récupère en faisant la somme de seulement 2 coefficients, $\tilde{c}_{(12)}$ et $\tilde{c}_{(13)}$. Plus généralement, on montre facilement que la composante d'ordre 2 dans la probabilité $\mathbb{P}[a_1 \prec ... \prec a_k]$ est donnée par :

$$\frac{1}{(k-1)!}\left(\sum_{i=1}^{k-1}(-1)^{\mathbb{I}\{a_i > a_{i+1}\}}\mathbb{P}[a_i \prec a_{i+1}] + \frac{\mathrm{as}(\pi) - \mathrm{ds}(\pi)}{2}\right),$$

où $\pi = a_1 \prec ... \prec a_k$, $\mathrm{as}(\pi) = \sum_{i=1}^{k-1} \mathbb{I}\{a_i < a_{i+1}\}$ est le nombre de *montées* (*ascents*) de $\pi$, et $\mathrm{ds}(\pi) = \sum_{i=1}^{k-1} \mathbb{I}\{a_i > a_{i+1}\}$ est le nombre de *descentes* (*descents*) de $\pi$. Dans la première construction, cette information est répartie sur tous les coefficients $c_{(i\,j)}$ tels que $\{i,j\} \cap \{a_1, ..., a_k\} \neq \emptyset$, soit $k(n-k)+1$ coefficients, alors qu'elle est répartie sur les coefficients $c_{(a_1\,a_2)}, ..., c_{(a_{k-1}\,a_k)}$ (soit $k-1$ coefficients) dans la deuxième.

Un autre point intéressant est que l'information d'ordre 2 qui intervient dans $\mathbb{P}[a_1 \prec ... \prec a_k]$ n'est donc contenue que dans les probabilités des classements sur les paires adjacentes dans $a_1 \prec ... \prec a_k$. D'ailleurs, on peut montrer facilement que la fonction $\mathbb{1}_{|\pi(i)-\pi(j)|=2}$ est d'ordre 3, *i.e* que sa projection sur $W^2$ est nulle, ou de manière équivalente que ses marginales d'ordre 2 sont toutes uniformes. Plus généralement, la fonction $\mathbb{1}_{|\pi(i)-\pi(j)|=k}$ est d'ordre $k-1$, même si elle ne fait intervenir que 2 objets. En fait elle est d'ordre 2 dans la décomposition de Fourier, puisqu'on a

$$\mathbb{1}_{|\pi(i)-\pi(j)|=k} = \sum_{\substack{1\leq i'\neq j'\leq n \\ |i'-j'|=k}} \mathbb{1}_{\{\pi(i)=i', \pi(j)=j'\}},$$

mais elle est d'ordre $k$ dans la notre (ou celle de Reiner). Une interprétation (avec les mains) que je propose est que si on ne stocke que des informations relatives, alors pour savoir que $\pi(i) - \pi(j)| = k$, on doit parcourir toute la liste des objets entre $i$ et $j$, soit $k+1$ éléments, alors que si on stocke des informations absolues (les rangs), il suffit de regarder le rang de $i$ et le rang de $j$, soit 2 éléments.

**Conséquences en pratique.** On considère le problème général de l'estimation, que l'on a formulé comme un problème inverse. On suppose donc qu'on a une probabilité $p$ sur $\mathfrak{S}_n$ que l'on observe à travers certaines marginales $p_A$ (où les sous-ensembles $A$ sont dans le support de la mesure $\mu$) et on veut estimer les probabilités des rankings sur un sous-ensemble $B$. On veut donc récupérer l'information contenue dans les observations, la synthétiser (pour réduire la variance ou la débruiter), et la transférer sur $B$. Supposons d'abord que l'on connaisse parfaitement les marginales observées (pas de bruit). Alors on peut décomposer aussi bien sur la première ou sur la deuxième base, l'information transférée sera la même. Or, par définition de la deuxième construction, l'information observée qui aura une influence sur $B$ est contenue dans les espaces $\tilde{W}_A^{|A|}$ pour $A \in \mathrm{supp}(\mu)$ et $A \subset B$. Autrement dit, on ne peut estimer des rankings sur $B$ qu'avec de l'information d'ordre inférieure, et strictement inférieure si $B$ n'a pas été observée. En effet, il n'y aurait pas de sens à ce que l'information d'ordre 3 relative aux classements sur $\{1,2,3\}$ soit utile pour prédire quelque chose sur $\{1,2,4\}$ puisque justement elle est relative aux interactions entre 1, 2 et 3 et pas seulement 1 et 2.

Cela pose a priori un problème si $|B| = 2$, puisque avec la deuxième construction on ne peut alors rien dire sur $B$. Et pourtant, on aurait envie de dire que si on observe $1 \prec 2$ et $2 \prec 3$ avec grande probabilité, alors on devrait avoir $1 \prec 3$ avec grande probabilité. Mais cette intuition

repose sur l'idée que la relation de transitivité devrait être "préservée" par la probabilité. Or en fait ce n'est pas nécessairement le cas, et pour le savoir, il faut appliquer la décomposition de Hodge de l'article *Statistical ranking and combinatorial hodge theory* qui décompose une probabilité $p$ de notre espace $V^2$ sur trois composantes :

- une composante acyclique qui préserve la transitivité (correspond au gradient),
- une composante localement cyclique qui casse la transitivté locale (correspond au rotationnel),
- une composante localement acyclique mais globalement cyclique (correspond au laplacien).

En l'occurrence, si les composantes acycliques sont fortes, la transitivité n'est pas nécessairement conservée. Par exemple, la probabilité

$$p = \frac{1}{3}\left(\delta_{123} + \delta_{231} + \delta_{312}\right)$$

est telle que $\mathbb{P}[1 \prec 2] = 2/3$ et $\mathbb{P}[2 \prec 3] = 2/3$ mais $\mathbb{P}[1 \prec 3] = 1/3$. Bref, pour prédire des comparaisons par paires, il faut faire une autre décomposition de $V^2$. Plus généralement, si on veut prédire sur un ensemble $B$ pour lequel on n'a pas pu inférer l'information d'ordre 2, il faut encore décomposer, sinon la deuxième décomposition est suffisante.

Enfin, d'un point de vue computationnel, si on observe une marginale sur $A$ avec $|A| = k$, sa décomposition dans la première base va potentiellement impliquer tous les espaces $W_{A'}^{|A'|}$ avec $|A'| \leq k$ et $A' \cap A \neq 0$. Ce qui fait potentiellement $O(n^{k-1})$ espaces, et donc $O(n^{k-1}(7/2)^k)$ coefficients. Dans le cas de la deuxième construction, les espaces impliqués sont les $\tilde{W}_{A'}^{|A'|}$ pour $A' \subset A$, ce qui revient à calculer la transformée en ondelettes de la marginale, donc avec complexité $O((5/2)^k)$. Grâce à sa forte localisation, la deuxième construction permet donc de s'affranchir d'une complexité en $n$, et de calculer localement les transformées en ondelettes de chaque marginale observée.

# Bibliography


Abbasnejad, E., Sanner, S., Bonilla, E. V., and Poupart, P. (2013). Learning community-based preferences via dirichlet process mixtures of gaussian processes. In *IJCAI*.

Adamowicz, W., Louviere, J., and Williams, M. (1994). Combining revealed and stated preference methods for valuing environmental amenities. *Journal of environmental economics and management*, 26(3):271–292.

Agarwal, S. (2006). Ranking on graph data. In *Proceedings of the 23rd international conference on Machine learning*, pages 25–32. ACM.

Agrawal, R., Rantzau, R., and Terzi, E. (2006). Context-sensitive ranking. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pages 383–394. ACM.

Aguiar, M. and Lauve, A. (2011). Lagrange's Theorem for Hopf Monoids in Species. *ArXiv e-prints*.

Ailon, N. (2008). Reconciling real scores with binary comparisons: A new logistic based model for ranking. In *Advances in Neural Information Processing Systems 21*, pages 25–32.

Ailon, N. (2010). Aggregation of partial rankings, p-ratings and top-m lists. *Algorithmica*, 57(2):284–300.

Ailon, N. (2012). An active learning algorithm for ranking from pairwise preferences with an almost optimal query complexity. *Journal of Machine Learning Research*, 13(1):137–164.

Ailon, N. (2014). Improved bounds for online learning over the permutahedron and other ranking polytopes. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, AISTATS 2014, Reykjavik, Iceland, April 22-25, 2014*, pages 29–37.

Ailon, N., Charikar, M., and Newman, A. (2008). Aggregating inconsistent information: ranking and clustering. *Journal of the ACM (JACM)*, 55(5):23.

Ailon, N., Hatano, K., and Takimoto, E. (2014). Bandit online optimization over the permutahedron. In *Algorithmic Learning Theory*, pages 215–229. Springer.

Ailon, N. and Mohri, M. (2010). Preference-based learning to rank. *Machine Learning*, 80(2-3):189–211.

Akritidis, L., Katsaros, D., and Bozanis, P. (2011). Effective rank aggregation for metasearching. *Journal of Systems and Software*, 84(1):130–143.

Akrour, R., Schoenauer, M., and Sebag, M. (2011). Preference-based policy learning. In *Machine learning and knowledge discovery in databases*, pages 12–27. Springer.

Alberto Maydeu-Olivares, A. and Hernández, A. (2007). Identification and small sample estimation of thurstone's unrestricted model for paired comparisons data. *Multivariate Behavioral Research*, 42(2):323–347.

Aldous, D. and Diaconis, P. (1986). Shuffling cards and stopping times. *American Mathematical Monthly*, pages 333–348.

Aledo, J. A., Gámez, J. A., and Molina, D. (2013). Tackling the rank aggregation problem with evolutionary algorithms. *Applied Mathematics and Computation*, 222:632–644.

Ali, A. and Meila, M. (2012). Experiments with kemeny ranking: What works when? *Mathematical Social Sciences*, 64(1):28 – 40.

Ali, M. M. (1998). Probability models on horse-race outcomes. *Journal of Applied Statistics*, 25(2):221–229.

Alvo, M. and Yu, P. (2014). *Statistical Methods for Ranking Data*. Springer.

Ammar, A. and Shah, D. (2012). Efficient rank aggregation using partial data. In *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '12, pages 355–366.

Arrow, K. (1951). Social choice and individual values.

Arrow, K. J. (1950). A difficulty in the concept of social welfare. *Journal of Political Economy*, 58(4):328–346.

Aslam, J. A. and Montague, M. (2001). Models for metasearch. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 276–284. ACM.

Audibert, J.-Y. and Tsybakov, A. (2007). Fast learning rates for plug-in classifiers. *Annals of statistics*, 35(2):608–633.

Awasthi, P., Blum, A., Sheffet, O., and Vijayaraghavan, A. (2014). Learning mixtures of ranking models. In *Advances in Neural Information Processing Systems 27*, pages 2609–2617.

Babington Smith, B. (1950). Discussion of professor ross's paper. *Journal of the Royal Statistical Society B*, 12(1):41–59.

Bachmaier, C., Brandenburg, F. J., Gleißner, A., and Hofmeier, A. (2013). On maximum rank aggregation problems. In *Combinatorial Algorithms*, pages 14–27. Springer.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern information retrieval*, volume 463. ACM press New York.

Balakrishnan, S. and Chopra, S. (2012). Collaborative ranking. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 143–152. ACM.

Balinski, M. L. and Laraki, R. (2010). *Majority judgment: measuring, ranking, and electing*. MIT press.

Baltrunas, L., Makcinskas, T., and Ricci, F. (2010). Group recommendations with rank aggregation and collaborative filtering. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 119–126. ACM.

Barg, A. and Mazumdar, A. (2010). Codes in permutations and error correction for rank modulation. *Information Theory, IEEE Transactions on*, 56(7):3158–3165.

Bargagliotti, A. E. (2009). Aggregation and decision making using ranked data. *Mathematical Social Sciences*, 58(3):354–366.

Bargagliotti, A. E. and Saari, D. G. (2010). Symmetry of nonparametric statistical tests on three samples. *Journal of Mathematics and Statistics*, 6(4):395–408.

Barjasteh, I., Forsati, R., Esfahanian, A.-H., and Radha, H. (2015). Semi-supervised collaborative ranking with push at top. *arXiv preprint*.

Barrow, D., Drayer, I., Elliott, P., Gaut, G., and Osting, B. (2013). Ranking rankings: an empirical comparison of the predictive power of sports ranking methods. *Journal of Quantitative Analysis in Sports*, 9(2):187–202.

Barthélémy, J. and Montjardet, B. (1981). The median procedure in cluster analysis and social choice theory. *Mathematical Social Sciences*, 1:235–267.

Bartholdi, J. J., Tovey, C. A., and Trick, M. A. (1989). The computational difficulty of manipulating an election. *Social Choice and Welfare*, 6:227–241.

Bartholdi, J. J., Tovey, C. A., and Trick, M. A. (1992). How hard is it to control an election? *Mathematical and Computer Modelling*, 16(8-9):27–40.

Barvinok, A. I. and Vershik, A. M. (1988). Methods of representations theory in combinatorial optimization problems. *Izv. Akad. Nauk SSSR Tekhn. Kibernet*, 205(6):64–71.

Basha, T., Moses, Y., and Avidan, S. (2012). Photo sequencing. In *Computer Vision–ECCV 2012*, pages 654–667. Springer.

Batchelder, W. H. and Bershad, N. J. (1979). The statistical analysis of a thurstonian model for rating chess players. *Journal of Mathematical Psychology*, 19(1):39–60.

Bayer, D. and Diaconis, P. (1992). Trailing the dovetail shuffle to its lair. *The Annals of Applied Probability*, pages 294–313.

Bedo, J. and Ong, C. S. (2014). Multivariate spearman's rho for aggregating ranks using copulas. *arXiv preprint*.

Behrisch, M., Davey, J., Simon, S., Schreck, T., Keim, D., and Kohlhammer, J. (2013). Visual Comparison of Orderings and Rankings. In *EuroVis Workshop on Visual Analytics*. The Eurographics Association.

Ben-Akiva, M. and Bierlaire, M. (1999). Discrete choice methods and their applications to short term travel decisions. In *Handbook of transportation science*, pages 5–33. Springer.

Ben-Akiva, M. E. and Lerman, S. R. (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press series in transportation studies. MIT Press.

Bennett, P. N., Chickering, D. M., and Mityagin, A. (2009). Learning consensus opinion: mining data from a labeling game. In *Proceedings of the 18th international conference on World wide web*, pages 121–130. ACM.

Benter, W. (1994). Computer-based horse race handicapping and wagering systems: A report. *Efficiency of racetrack betting markets*, pages 183–198.

Berbeglia, G. (2016). Discrete choice models based on random walks. *Operations Research Letters*, 44(2):234 – 237.

Berry, S., Levinsohn, J., and Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, pages 841–890.

Betzler, N., Bredereck, R., and Niedermeier, R. (2014). Theoretical and empirical evaluation of data reduction for exact kemeny rank aggregation. *Autonomous Agents and Multi-Agent Systems*, 28(5):721–748.

Betzler, N., Fellows, M. R., Guo, J., Niedermeier, R., and Rosamond, F. A. (2008). Fixed-parameter algorithms for kemeny scores. In *Algorithmic Aspects in Information and Management*, pages 60–71. Springer.

Betzler, N., Fellows, M. R., Guo, J., Niedermeier, R., and Rosamond, F. A. (2009). How similarity helps to efficiently compute kemeny rankings. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 657–664. International Foundation for Autonomous Agents and Multiagent Systems.

Biernacki, C. and Jacques, J. (2013). A generative model for rank data based on insertion sort algorithm. *Computational Statistics & Data Analysis*, 58:162–176.

Björner, A. and Wachs, M. L. (1983). On lexicographically shellable posets. *Trans. Amer. Math. Soc.*, 277:323–341.

Blanchet, J. H., Gallego, G., and Goyal, V. (2013). A markov chain approximation to choice modeling. In *EC*, pages 103–104.

Blin, G., Crochemore, M., Hamel, S., and Vialette, S. (2011). Median of an odd number of permutations. *Pure Mathematics and Applications*, 21(2):161–175.

Block, H. D. and Marschak, J. (1960). Random orderings and stochastic theories of responses. *Contributions to probability and statistics*, 2:97–132.

Bock, R. D. and Jones, J. V. (1968). The measurement and prediction of judgment and choice.

Böckenholt, U. (1992). Thurstonian representation for partial ranking data. *British Journal of Mathematical and Statistical Psychology*, 45(1):31–49.

Böckenholt, U. (2006). Thurstonian-based analyses: Past, present, and future utilities. *Psychometrika*, 71(4):615–629.

Borda, J. C. (1781). Mémoire sur les élections au scrutin.

Boulesteix, A.-L. and Slawski, M. (2009). Stability and aggregation of ranked gene lists. *Briefings in bioinformatics*, 10(5):556–568.

Bradley, R. A. (1976). Science, statistics, and paired comparisons. *Biometrics*, pages 213–239.

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Brandenburg, F.-J., Gleißner, A., and Hofmeier, A. (2012). Comparing and aggregating partial orders with kendall tau distances. In *WALCOM*, pages 88–99.

Brandt, F., Brill, M., Hemaspaandra, E., and Hemaspaandra, L. A. (2015). Bypassing combinatorial protections: Polynomial-time algorithms for single-peaked electorates. *Journal of Artificial Intelligence Research*, pages 439–496.

Braverman, M. and Mossel, E. (2008). Noisy sorting without resampling. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '08, pages 268–276.

Breitling, R., Armengaud, P., Amtmann, A., and Herzyk, P. (2004). Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS letters*, 573(1):83–92.

Brinker, K. and Hüllermeier, E. (2007). Case-based multilabel ranking. In *IJCAI*, pages 702–707.

Brochu, E., de Freitas, N., and Ghosh, A. (2008). Active preference learning with discrete choice data. In *Advances in neural information processing systems*, pages 409–416.

Brown, T. C., Nannini, D., Gorter, R. B., Bell, P. A., and Peterson, G. L. (2002). Judged seriousness of environmental losses: reliability and cause of loss. *Ecological Economics*, 42(3):479–491.

Buhlmann, H. and Huber, P. J. (1963). Pairwise comparison and ranking in tournaments. *The Annals of Mathematical Statistics*, 34(2):501–510.

Buhyoff, G. J. and Leuschner, W. A. (1978). Estimating psychological disutility from damaged forest stands. *Forest Science*, 24(3):424–432.

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., and Hullender, G. (2005). Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 89–96. ACM International Conference Proceeding Series **119**.

Burges, C. J., Ragno, R., and Le, Q. V. (2006). Learning to rank with nonsmooth cost functions. In Schölkopf, B., Platt, J. C., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*, pages 193–200. MIT Press.

Busa-Fekete, R., Hüllermeier, E., and Szörényi, B. (2014a). Preference-based rank elicitation using statistical models: The case of mallows. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1071–1079.

Busa-Fekete, R., Szarvas, G., Elteto, T., and Kégl, B. (2012). An apple-to-apple comparison of learning-to-rank algorithms in terms of normalized discounted cumulative gain. In *20th European Conference on Artificial Intelligence (ECAI 2012): Preference Learning: Problems and Applications in AI Workshop*, volume 242. Ios Press.

Busa-Fekete, R., Szorenyi, B., Cheng, W., Weng, P., and Hüllermeier, E. (2013). Top-k selection based on adaptive sampling of noisy preferences. In *Proceedings of The 30th International Conference on Machine Learning*, pages 1094–1102.

Busa-Fekete, R., Szörényi, B., and Hüllermeier, E. (2014b). Pac rank elicitation through adaptive sampling of stochastic pairwise preferences. In *28th AAAI Conference on Artificial Intelligence (AAAI-14)*.

Busse, L. M., Orbanz, P., and Buhmann, J. M. (2007). Cluster analysis of heterogeneous rank data. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pages 113–120.

Cao, Z., Qin, T., Liu, T.-Y., Tsai, M.-F., and Li, H. (2007). Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136. ACM.

Caragiannis, I., Procaccia, A. D., and Shah, N. (2013). When do noisy votes reveal the truth? In *Proceedings of the fourteenth ACM conference on Electronic commerce*, pages 143–160. ACM.

Caragiannis, I., Procaccia, A. D., and Shah, N. (2014). Modal ranking: A uniquely robust voting rule. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Carlsson, C. and Fullér, R. (1996). Fuzzy multiple criteria decision making: Recent developments. *Fuzzy sets and systems*, 78(2):139–153.

Caron, F. and Doucet, A. (2012). Efficient bayesian inference for generalized bradley-terry models. *Journal of Computational and Graphical Statistics*, 21(1):174–196.

Caron, F. and Teh, Y. W. (2012). Bayesian nonparametric models for ranked data. In *Advances in Neural Information Processing Systems*, pages 1520–1528.

Caron, F., Teh, Y. W., and Murphy, T. B. (2014). Bayesian nonparametric plackett–luce models for the analysis of preferences for college degree programmes. *The Annals of Applied Statistics*, 8(2):1145–1181.

Cattelan, M. (2012). Models for paired comparison data: A review with emphasis on dependent data. *Statist. Sci.*, 27(3):412–433.

Cattelan, M., Varin, C., and Firth, D. (2013). Dynamic bradley–terry modelling of sports tournaments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(1):135–150.

Ceberio, J., Irurozki, E., Mendiburu, A., and Lozano, J. A. (2014). Extending distance-based ranking models in estimation of distribution algorithms. In *Evolutionary Computation (CEC), 2014 IEEE Congress on*, pages 2459–2466. IEEE.

Ceccherini-Silberstein, T., Scarabotti, F., and Tolli, F. (2010). *Representation theory of the symmetric groups: the Okounkov-Vershik approach, character formulas, and partition algebras*, volume 121. Cambridge University Press.

Chandra, A. and Roy, S. (2013). On removing condorcet effects from pairwise election tallies. *Social Choice and Welfare*, 40(4):1143–1158.

Chao, M. and Strawderman, W. (1972). Negative moments of positive random variables. *Journal of the American Statistical Society*, 67:429–431.

Chapelle, O. and Chang, Y. (2011). Yahoo! learning to rank challenge overview. In *Yahoo! Learning to Rank Challenge*, pages 1–24.

Chaudhuri, S. and Tewari, A. (2015). Online ranking with top-1 feedback. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 129–137.

Chechik, G., Sharma, V., Shalit, U., and Bengio, S. (2010). Large scale online learning of image similarity through ranking. *The Journal of Machine Learning Research*, 11:1109–1135.

Chen, K. T., Fox, R. H., and Lyndon, R. C. (1958). Free differential calculus, iv. the quotient groups of the lower central series. *Annals of Mathematics*, pages 81–95.

Chen, X., Bennett, P. N., Collins-Thompson, K., and Horvitz, E. (2013). Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 193–202.

Chen, Y. and Suh, C. (2015). Spectral mle: Top-k rank aggregation from pairwise comparisons. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 371–380.

Cheng, W., Dembczyński, K., and Hüllermeier, E. (2010). Label ranking methods based on the Plackett-Luce model. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 215–222.

Cheng, W., Fürnkranz, J., Hüllermeier, E., and Park, S.-H. (2011). Preference-based policy iteration: Leveraging preference learning for reinforcement learning. In *Machine learning and knowledge discovery in databases*, pages 312–327. Springer.

Cheng, W., Hühn, J., and Hüllermeier, E. (2009). Decision tree and instance-based learning for label ranking. In *Proceedings of the 26th International Conference on Machine Learning (ICML-09)*, pages 161–168.

Cheng, W., Hüllermeier, E., Waegeman, W., and Welker, V. (2012). Label ranking with partial abstention based on thresholded probabilistic models. In *Advances in Neural Information Processing Systems 25*, pages 2501–2509.

Chierichetti, F., Dasgupta, A., Kumar, R., and Lattanzi, S. (2015). On learning mixture models for permutations. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, pages 85–92. ACM.

Chin, F. Y. L., Deng, X., Fang, Q., and Zhu, S. (2004). Approximate and dynamic rank aggregation. *Theoretical computer science*, 325(3):409–424.

Chu, M. T. (1998). On the optimal consistent approximation to pairwise comparison matrices. *Linear Algebra and Its Applications*, 272(1):155–168.

Chu, W. and Ghahramani, Z. (2005a). Gaussian processes for ordinal regression. In *Journal of Machine Learning Research*, pages 1019–1041.

Chu, W. and Ghahramani, Z. (2005b). Preference learning with gaussian processes. In *Proceedings of the 22nd International Conference on Machine learning*, pages 137–144. ACM.

Chung, L. and Marden, J. I. (1993). Extensions of mallows' $\phi$ model. In Fligner, M. A. and Verducci, J. S., editors, *Probability Models and Statistical Analyses for Ranking Data*, volume 80 of *Lecture Notes in Statistics*, pages 108–139. Springer New York.

Clémençon, S., Lugosi, G., and Vayatis, N. (2008). Ranking and empirical risk minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874.

Clémençon, S. and Robbiano, S. (2011). Minimax learning rates for bipartite ranking and plug-in rules. In *Proceedings of the International Conference in Machine Learning, ICML'11*.

Clémençon, S. and Vayatis, N. (2007). Ranking the best instances. *Journal of Machine Learning Research*, 8:2671–2699.

Clémençon, S., Gaudel, R., and Jakubowicz, J. (2011). Clustering rankings in the fourier domain. In *Machine Learning and Knowledge Discovery in Databases*, pages 343–358. Springer.

Clémençon, S., Jakubowicz, J., and Sibony, E. (2014). Multiresolution analysis of incomplete rankings. *arXiv preprint*.

Cohen, M. and Falmagne, J.-C. (1990). Random utility representation of binary choice probabilities: a new class of necessary conditions. *Journal of Mathematical Psychology*, 34(1):88–94.

Cohen, W. W., Schapire, R. E., and Singer, Y. (1999). Learning to order things. *Journal of Artificial Intelligence Research*, 10(1):243–270.

Coifman, R. and Maggioni, M. (2006). Diffusion wavelets. *Applied and Computational Harmonic Analysis*, 21:53–94.

Condorcet, N. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. L'imprimerie royale, Paris.

Conitzer, V., Davenport, A., and Kalagnanam, J. (2006). Improved bounds for computing kemeny rankings. In *AAAI*, volume 6, pages 620–626.

Conitzer, V., Rognlie, M., and Xia, L. (2009). Preference functions that score rankings and maximum likelihood estimation. In *IJCAI*, volume 9, pages 109–115.

Conitzer, V. and Sandholm, T. (2005). Common voting rules as maximum likelihood estimators. In *Proceedings of the Twenty-First Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-05)*, pages 145–152, Arlington, Virginia. AUAI Press.

Conitzer, V., Sandholm, T., and Lang, J. (2007). When are elections with few candidates hard to manipulate? *Journal of the ACM (JACM)*, 54(3):14.

Cooper, W. S., Gey, F. C., and Dabney, D. P. (1992). Probabilistic retrieval based on staged logistic regression. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 198–210. ACM.

Copeland, A. H. (1951). A reasonable social welfare function. In *Seminar on applications of mathematics to social sciences, University of Michigan*.

Coppersmith, D., Fleischer, L., and Rudra, A. (2006). Ordering by weighted number of wins gives a good ranking for weighted tournaments. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithm*, SODA '06, pages 776–782.

Cornaz, D., Galand, L., and Spanjaard, O. (2013). Kemeny elections with bounded single-peaked or single-crossing width. In *IJCAI*, volume 13, pages 76–82. Citeseer.

Cossock, D. and Zhang, T. (2006). Subset ranking using regression. In *Learning theory*, pages 605–619. Springer.

Cossock, D. and Zhang, T. (2008). Statistical analysis of bayes optimal subset ranking. *Information Theory, IEEE Transactions on*, 54(11):5140–5154.

Crammer, K. and Singer, Y. (2001). Pranking with ranking. In *Advances in Neural Information Processing Systems*, pages 641–647.

Crammer, K. and Singer, Y. (2003). A family of additive online algorithms for category ranking. *The Journal of Machine Learning Research*, 3:1025–1058.

Crisman, K.-D. (2014). The Borda count, the Kemeny rule, and the permutahedron. *Contemporary Mathematics*, 624.

Critchlow, D. E. (1985). *Metric Methods for Analyzing Partially Ranked Data*, volume 34 of *Lecture Notes in Statistics*. Springer.

Critchlow, D. E., Fligner, M. A., and Verducci, J. S. (1991). Probability models on rankings. *Journal of Mathematical Psychology*, 35(3):294 – 318.

Croon, M. A. (1989). Latent class models for the analysis of rankings. In Geert de Soete, H. F. and Klauer, K. C., editors, *New Developments in Psychological Choice Modeling*, volume 60 of *Advances in Psychology*, pages 99 – 121. North-Holland.

Csiszár, V. (2008). Conditional independence relations and log-linear models for random matchings. *Acta Mathematica Hungarica*, 122(1-2):131–152.

Csiszár, V. (2009a). Markov bases of conditional independence models for permutations. *Kybernetika*, 45:249–260.

Csiszár, V. (2009b). On l-decomposability of random orderings. *Journal of Mathematical Psychology*, 53(4):294 – 297.

Cucuringu, M. (2015). Sync-rank: Robust ranking, constrained ranking and rank aggregation via eigenvector and semidefinite programming synchronization. *arXiv preprint*.

Dalal, O., Sengemedu, S. H., and Sanyal, S. (2012). Multi-objective ranking of comments on web. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 419–428.

Dangauthier, P., Herbrich, R., Minka, T., and Graepel, T. (2007). Trueskill through time: Revisiting the history of chess. In *Advances in Neural Information Processing Systems*, pages 337–344.

Daugherty, Z., Eustis, A. K., Minton, G., and Orrison, M. E. (2009). Voting, the symmetric group, and representation theory. *The American Mathematical Monthly*, 116(8):667–687.

Davenport, A. and Kalagnanam, J. (2004). A computational study of the kemeny rule for preference aggregation. In *AAAI*, volume 4, pages 697–702.

David, H. A. (1963). *The method of paired comparisons*. Griffin's statistical monographs & courses. Hafner Pub. Co.

Davidson, R. R. (1970). On extending the bradley-terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65(329):317–328.

Davidson, R. R. and Farquhar, P. H. (1976). A bibliography on the method of paired comparisons. *Biometrics*, 32:241–252.

Debreu, G. (1960). Review of r. d. luce, individual choice behavior: A theoretical analysis. *American Economic Review*, 50:186–188.

DeConde, R. P., Hawley, S., Falcon, S., Clegg, N., Knudsen, B., and Etzioni, R. (2006). Combining results of microarray experiments: a rank aggregation approach. *Statistical Applications in Genetics and Molecular Biology*, 5(1).

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391.

Dekel, O., Singer, Y., and Manning, C. D. (2003). Log-linear models for label ranking. In *Advances in neural information processing systems*, page None.

D'Elia, A. (2000). A shifted binomial model for rankings. In *Statistical Modelling, XV International Workshop on Statistical Modelling, Servicio Editorial de la Universidad del Pais Vasco*, pages 412–416.

D'Elia, A. (2003). Modelling ranks using the inverse hypergeometric distribution. *Statistical Modelling*, 3(1):65–78.

Deng, K., Han, S., Li, K. J., and Liu, J. S. (2014). Bayesian aggregation of order-based rank data. *Journal of the American Statistical Association*, 109(507):1023–1039.

Desarkar, M. S., Sarkar, S., and Mitra, M. (2016). Preference relations based unsupervised rank aggregation for metasearch. *Expert Systems with Applications*, 49:86 – 98.

Désir, A., Goyal, V., Segev, D., and Ye, C. (2015). Capacity constrained assortment optimization under the markov chain based choice model. *Operations Research, Forthcoming*.

Destercke, S. (2013). A pairwise label ranking method with imprecise scores and partial predictions. In *Machine Learning and Knowledge Discovery in Databases*, pages 112–127. Springer.

DeVore, R. A. (1998). Nonlinear approximation. *Acta numerica*, 7:51–150.

Deza, M. and Huang, T. (1998). Metrics on permutations, a survey.

Diaconis, P. (1988). *Group representations in probability and statistics*. Institute of Mathematical Statistics Lecture Notes - Monograph Series. Institute of Mathematical Statistics, Hayward, CA.

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

Diaconis, P. and Eriksson, N. (2006). Markov bases for noncommutative fourier analysis of ranked data. *Journal of Symbolic Computation*, 41(2):182 – 195.

Diaconis, P., Graham, R., and Holmes, S. P. (2001). Statistical problems involving permutations with restricted positions. *Lecture Notes-Monograph Series*, pages 195–222.

Diaconis, P. and Graham, R. L. (1977). Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 262–268.

Diaconis, P., Pang, C. Y. A., and Ram, A. (2014). Hopf algebras and markov chains: two examples and a theory. *Journal of Algebraic Combinatorics*, 39(3):527–585.

Diaconis, P. and Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. *The Annals of Statistics*, 26(1):363–397.

Ding, W., Ishwar, P., and Saligrama, V. (2015a). Learning mixed membership mallows models from pairwise comparisons. *arXiv preprint*.

Ding, W., Ishwar, P., and Saligrama, V. (2015b). A topic modeling approach to ranking. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*.

Dittman, D. J., Khoshgoftaar, T. M., Wald, R., and Napolitano, A. (2013). Classification performance of rank aggregation techniques for ensemble gene selection. In *The Twenty-Sixth International FLAIRS Conference*.

Doignon, J.-P., Pekeč, A., and Regenwetter, M. (2004). The repeated insertion model for rankings: Missing link between two subset choice models. *Psychometrika*, 69(1):33–54.

Donoho, D. (1993). Unconditional bases are optimal bases for data compression and for statistical estimation. *Appl. Comput. Harm. Analysis*, 1:100–115.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1996). Density estimation by wavelet thresholding. *Annals of Statistics*, 24(2):508–539.

Dopazo, E. and González-Pachón, J. (2003). Consistency-driven approximation of a pairwise comparison matrix. *Kybernetika*, 39(5):561–568.

Duchi, J. C., Mackey, L., and Jordan, M. I. (2013). The asymptotics of ranking algorithms. *The Annals of Statistics*, 41(5):2292–2323.

Dwork, C., Kumar, R., Naor, M., and Sivakumar, D. (2001). Rank aggregation methods for the Web. In *Proceedings of the 10th International WWW conference*, pages 613–622.

Dyer, J. S., Fishburn, P. C., Steuer, R. E., Wallenius, J., and Zionts, S. (1992). Multiple criteria decision making, multiattribute utility theory the next ten years. *Management science*, 38(5):645–654.

Elisseeff, A. and Weston, J. (2001). A kernel method for multi-labelled classification. In *Advances in neural information processing systems*, pages 681–687.

Elkind, E., Faliszewski, P., and Slinko, A. (2015). Distance rationalization of voting rules. *Social Choice and Welfare*, 45(2):345–377.

Elo, A. E. (1978). *The rating of chessplayers, past and present*. Arco Pub.

Erdélyi, G., Piras, L., and Rothe, J. (2011). The complexity of voter partition in bucklin and fallback voting: Solving three open problems. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 837–844. International Foundation for Autonomous Agents and Multiagent Systems.

Eriksson, B. (2013). Learning to top-k search using pairwise comparisons. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pages 265–273.

Eugster, M. J. A., Leisch, F., and Strobl, C. (2014). (psycho-) analysis of benchmark experiments: A formal framework for investigating the relationship between data sets and learning algorithms. *Computational Statistics & Data Analysis*, 71:986–1000.

Fagin, R., Kumar, R., Mahdian, M., Sivakumar, D., and Vee, E. (2004). Comparing and aggregating rankings with ties. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 47–58. ACM.

Fagin, R., Kumar, R., Mahdian, M., Sivakumar, D., and Vee, E. (2006). Comparing partial rankings. *SIAM J. Discrete Mathematics*, 20(3):628–648.

Fagin, R., Kumar, R., and Sivakumar, D. (2003). Efficient similarity search and classification via rank aggregation. In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, pages 301–312. ACM.

Faliszewski, P., Hemaspaandra, E., Hemaspaandra, L. A., and Rothe, J. (2009). A richer understanding of the complexity of election systems. In *Fundamental Problems in Computing*, pages 375–406. Springer.

Faliszewski, P. and Procaccia, A. D. (2010). Ai's war on manipulation: Are we winning? *AI Magazine*, 31(4):53–64.

Farias, V., Jagabathula, S., and Shah, D. (2009). A data-driven approach to modeling choice. In *Advances in Neural Information Processing Systems*, pages 504–512.

Farias, V. F., Jagabathula, S., and Shah, D. (2013). A nonparametric approach to modeling choice with limited data. *Management Science*, 59(2):305–322.

Farmer, F. (1978). Cellular homology for posets. *Math. Japon*, 23:607–613.

Farnoud, F. and Milenkovic, O. (2014). An axiomatic approach to constructing distances for rank comparison and aggregation. *Information Theory, IEEE Transactions on*, 60(10):6417–6439.

Farnoud, F., Schwartz, M., and Bruck, J. (2014). Bounds for permutation rate-distortion. In *Information Theory (ISIT), 2014 IEEE International Symposium on*, pages 6–10. IEEE.

Fasola, S. and Sciandra, M. (2015). *Advances in Statistical Models for Data Analysis*, chapter New Flexible Probability Distributions for Ranking Data, pages 117–124. Springer International Publishing, Cham.

Fechner, G. T. (1860). *Elemente der Psychophysik*. Number 1 in Elemente der Psychophysik. Breitkopf und Härtel.

Feigin, P. D. and Alvo, M. (1986). Intergroup diversity and concordance for ranking data: An approach via metrics for permutations. *Ann. Statist.*, 14(2):691–707.

Feigin, P. D. and Cohen, A. (1978). On a model for concordance between judges. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 203–213.

Feng, J., Fang, Q., and Ng, W. (2008). Discovering bucket orders from full rankings. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 55–66. ACM.

Fernau, H., Fomin, F. V., Lokshtanov, D., Mnich, M., Philip, G., and Saurabh, S. (2010). Ranking and drawing in subexponential time. In *Combinatorial Algorithms*, pages 337–348. Springer.

Figueira, J., Greco, S., and Ehrgott, M. (2005). *Multiple criteria decision analysis: state of the art surveys*, volume 78. Springer Science & Business Media.

Fishburn, P. C. (1973). Binary choice probabilities: on the varieties of stochastic transitivity. *Journal of Mathematical psychology*, 10(4):327–352.

Fishburn, P. C. (1977). Condorcet social choice functions. *SIAM Journal on applied Mathematics*, 33(3):469–489.

Fishburn, P. C. (1992). Induced binary probabilities and the linear ordering polytope: A status report. *Mathematical Social Sciences*, 23(1):67–80.

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

Fligner, M. A. and Verducci, J. S. (1988). Multistage ranking models. *Journal of the American Statistical Association*, 83(403):892–901.

Fogel, F., Jenatton, R., Bach, F., and d'Aspremont, A. (2013). Convex relaxations for permutation problems. In *Advances in Neural Information Processing Systems*, pages 1016–1024.

Ford, L. R. (1957). Solution of a ranking problem from binary comparisons. *The American Mathematical Monthly*, 64(8):28–33.

Freund, D. and Williamson, D. P. (2015). Rank aggregation: New bounds for mcx. *CoRR*, abs/1510.00738.

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

Friedman, J. (1996). Another approach to polychotomous classification. Technical report, Technical report, Department of Statistics, Stanford University.

Fuhr, N. (1992). Probabilistic models in information retrieval. *The Computer Journal*, 35(3):243–255.

Fülöp, J. (2008). A method for approximating pairwise comparison matrices by consistent matrices. *Journal of Global Optimization*, 42(3):423–442.

Fulton, W. and Harris, J. (1991). *Representation theory*, volume 129. Springer Science & Business Media.

Fürnkranz, J. (2002). Round robin classification. *The Journal of Machine Learning Research*, 2:721–747.

Fürnkranz, J. and Hüllermeier, E. (2011). *Preference learning*. Springer.

Gallego, G., Ratliff, R., and Shebalov, S. (2014). A general attraction model and sales-based linear program for network revenue management under customer choice. *Operations Research*, 63(1):212–232.

Gavish, M., Nadler, B., and Coifman, R. R. (2010). Multiscale wavelets on trees, graphs and high dimensional data: theory and applications to semi supervised learning. In *International Conference on Machine Learning*, pages 567–574.

Geman, D., dÁvignon, C., Naiman, D. Q., and Winslow, R. L. (2004). Classifying gene expression profiles from pairwise mrna comparisons. *Statistical Applications in Genetics and Molecular Biology*, 3.

Georgescu-Roegen, N. (1958). Threshold in choice and the theory of demand. *Econometrica: Journal of the Econometric Society*, pages 157–168.

Ghahramani, Z. and Jordan, M. I. (1995). Learning from incomplete data. Technical report, Lab Memo No. 1509, CBCL Paper No. 108, MIT AI Lab.

Gibbard, A. (1973). Manipulation of voting schemes: a general result. *Econometrica: journal of the Econometric Society*, pages 587–601.

Gionis, A., Mannila, H., Puolamäki, K., and Ukkonen, A. (2006). Algorithms for discovering bucket orders from data. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 561–566. ACM.

Gleich, D. F. and Lim, L.-H. (2011). Rank aggregation via nuclear norm minimization. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 60–68.

Glickman, M. E. and Jensen, S. T. (2005). Adaptive paired comparison design. *Journal of statistical planning and inference*, 127(1):279–293.

Goldsmith, J., Lang, J., Mattei, N., and Perny, P. (2014). Voting with rank dependent scoring rules. In *AAAI*, pages 698–704.

Gormley, I. C. and Murphy, T. B. (2008). A mixture of experts model for rank data with applications in election studies. *The Annals of Applied Statistics*, 2(4):1452–1477.

Gormley, I. C. and Murphy, T. B. (2009). A grade of membership model for rank data. *Bayesian Analysis*, 4(2):265–295.

Gratzl, S., Lex, A., Gehlenborg, N., Pfister, H., and Streit, M. (2013). Lineup: Visual analysis of multi-attribute rankings. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2277–2286.

Grbovic, M., Djuric, N., and Vucetic, S. (2013). Multi-prototype label ranking with novel pairwise-to-total-rank aggregation. In *IJCAI*.

Greco, S., Matarazzo, B., and Slowinski, R. (2001). Rough sets theory for multicriteria decision analysis. *European journal of operational research*, 129(1):1–47.

Green, P. E., Krieger, A. M., and Wind, Y. (2001). Thirty years of conjoint analysis: Reflections and prospects. *Interfaces*, 31(3):S56–S73.

Green, P. E. and Rao, V. R. (1971). Conjoint measurement for quantifying judgmental data. *Journal of Marketing Research*, 8(3):355–363.

Green, P. E. and Srinivasan, V. (1978). Conjoint analysis in consumer research: Issues and outlook. *Journal of Consumer Research*, 5(2):103–123.

Green, P. E. and Srinivasan, V. (1990). Conjoint analysis in marketing: New developments with implications for research and practice. *Journal of Marketing*, 54(4):3–19.

Grötschel, M., Jünger, M., and Reinelt, G. (1985). Facets of the linear ordering polytope. *Mathematical Programming*, 33(1):43–60.

Guadagni, P. M. and Little, J. D. C. (1983). A logit model of brand choice calibrated on scanner data. *Marketing science*, 2(3):203–238.

Guilford, J. P. (1954). Psychometric methods.

Guiver, J. and Snelson, E. (2009). Bayesian inference for plackett-luce ranking models. In *ICML*.

Gul, F., Natenzon, P., and Pesendorfer, W. (2014). Random choice as behavioral optimization. *Econometrica*, 82(5):1873–1912.

Guo, S. and Sanner, S. (2010). Real-time multiattribute bayesian preference elicitation with pairwise comparison queries. In *International Conference on Artificial Intelligence and Statistics*, pages 289–296.

Guo, S., Sanner, S., and Bonilla, E. V. (2010). Gaussian process preference elicitation. In *Advances in Neural Information Processing Systems*, pages 262–270.

Hajek, B., Oh, S., and Xu, J. (2014). Minimax-optimal inference from partial rankings. In *Advances in Neural Information Processing Systems*, pages 1475–1483.

Hall, P. and Miller, H. (2010). Modeling the variability of rankings. *Ann. Statist.*, 38(5):2652–2677.

Hammond, D. K., Vandergheynst, P., and Gribonval, R. (2011). Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2):129 – 150.

Hastie, T. and Tibshirani, R. (1998). Classification by pairwise coupling. *The annals of statistics*, 26(2):451–471.

Haunsperger, D. B. and Saari, D. G. (1991). The lack of consistency for statistical decision procedures. *The American Statistician*, 45(3):252–255.

Hausman, J. and McFadden, D. (1984). Specification tests for the multinomial logit model. *Econometrica*, 52(5):1219–1240.

He, Z. and Yu, W. (2010). Stable feature selection for biomarker discovery. *Computational biology and chemistry*, 34(4):215–225.

Helmbold, D. P. and Warmuth, M. K. (2009). Learning permutations with exponential weights. *Journal of Machine Learning Research*, 10:1705–1736.

Helmi, A., Lumbroso, J., Martínez, C., and Viola, A. (2012). Data streams as random permutations: the distinct element problem. *DMTCS Proceedings*, (01):323–338.

Hemaspaandra, E., Hemaspaandra, L. A., and Rothe, J. (1997). Exact analysis of dodgson elections: Lewis carroll's 1876 voting system is complete for parallel access to np. *Journal of the ACM (JACM)*, 44(6):806–825.

Henery, R. J. (1981). Permutation probabilities as models for horse races. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 86–91.

Henery, R. J. (1992). An extension to the thurstone-mosteller model for chess. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 41(5):559–567.

Herbrich, R., Graepel, T., and Obermayer, K. (1999). Large margin rank boundaries for ordinal regression. *Advances in neural information processing systems*, pages 115–132.

Herbrich, R., Minka, T., and Graepel, T. (2006). Trueskill$^{\text{TM}}$: A bayesian skill rating system. In *Advances in Neural Information Processing Systems*, pages 569–576.

Hornik, K. and Meyer, D. (2007). *Deriving consensus rankings from benchmarking experiments*. Springer.

Houlsby, N., Huszar, F., Ghahramani, Z., and Hernández-lobato, J. M. (2012). Collaborative gaussian processes for preference learning. In *Advances in Neural Information Processing Systems*, pages 2096–2104.

Huang, J. (2011). *Probabilistic Reasoning and Learning on Permutations: Exploiting Structural Decompositions of the Symmetric Group*. PhD thesis, Carnegie Mellon University.

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

Huang, J. and Guestrin, C. (2012). Uncovering the riffled independence structure of ranked data. *Electronic Journal of Statistics*, 6:199–230.

Huang, J., Guestrin, C., and Guibas, L. (2007). Efficient inference for distributions on permutations. In *Advances in Neural Information Processing Systems 20*, pages 697–704.

Huang, J., Guestrin, C., and Guibas, L. (2009a). Fourier theoretic probabilistic inference over permutations. *JMLR*, 10:997–1070.

Huang, J., Guestrin, C., Jiang, X., and Guibas, L. J. (2009b). Exploiting probabilistic independence for permutations. In *International Conference on Artificial Intelligence and Statistics*, pages 248–255.

Huang, J., Kapoor, A., and Guestrin, C. (2012). Riffled independence for efficient inference with partial ranking. *Journal of Artificial Intelligence*, 44:491–532.

Huang, T.-K., Weng, R. C., and Lin, C.-J. (2006). Generalized bradley-terry models and multi-class probability estimates. *The Journal of Machine Learning Research*, 7:85–115.

Hudry, O. (2008). NP-hardness results for the aggregation of linear orders into median orders. *Ann. Oper. Res.*, 163:63–88.

Hüllermeier, E., Fürnkranz, J., Cheng, W., and Brinker, K. (2008). Label ranking by learning pairwise preferences. *Artificial Intelligence*, 172(16):1897–1916.

Hunter, D. R. (2004). MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics*, 32:384–406.

Ieong, S., Mishra, N., and Sheffet, O. (2012). Predicting preference flips in commerce search. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*.

Irurozki, E., Calvo, B., and Lozano, J. (2011). Learning probability distributions over permutations by means of Fourier coefficients. *Advances in Artificial Intelligence*, pages 186–191.

Jacques, J. and Biernacki, C. (2014). Model-based clustering for multivariate partial ranking data. *Journal of Statistical Planning and Inference*, 149:201–217.

Jagabathula, S. and Shah, D. (2008). Inferring rankings under constrained sensing. In *Advances in Neural Information Processing Systems*, pages 753–760.

Jagabathula, S. and Shah, D. (2011). Inferring Rankings Using Constrained Sensing. *IEEE Transactions on Information Theory*, 57(11):7288–7306.

James, G. and Kerber, A. (1981). The representation theory of the symmetric group. *Reading, Mass.*

Jamieson, K. G. and Nowak, R. (2011). Active ranking using pairwise comparisons. In *Advances in Neural Information Processing Systems 24*, pages 2240–2248.

Jang, M., Kim, S., Suh, C., and Oh, S. (2016). Top-$k$ ranking from pairwise comparisons: When spectral ranking is optimal. *arXiv preprint*.

Jech, T. (1983). The ranking of incomplete tournaments: A mathematician's guide to popular sports. *The American Mathematical Monthly*, 90(4):246–266.

Jiang, X., Huang, J., and Guibas, L. (2011a). Fourier-information duality in the identity management problem. In *Machine Learning and Knowledge Discovery in Databases*, pages 97–113. Springer.

Jiang, X., Lim, L.-H., Yao, Y., and Ye, Y. (2011b). Statistical ranking and combinatorial Hodge theory. *Math. Program.*, 127(1):203–244.

Jiang, X., Sun, J., and Guibas, L. (2014). A fourier-theoretic approach for inferring symmetries. *Computational Geometry*, 47(2):164–174.

Jiao, Y. and Vert, J.-P. (2015). The kendall and mallows kernels for permutations. In Blei, D. and Bach, F., editors, *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1935–1944. JMLR Workshop and Conference Proceedings.

Jin, R., Si, L., Zhai, C., and Callan, J. (2003). Collaborative filtering with decoupled models for preferences and ratings. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 309–316. ACM.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM.

Joe, H. and Verducci, J. S. (1993). *Probability Models and Statistical Analyses for Ranking Data*, chapter On the Babington Smith Class of Models for Rankings, pages 37–52. Springer New York, New York, NY.

Kakarala, R. (2011). A signal processing approach to Fourier analysis of ranking data: the importance of phase. *IEEE Transactions on Signal Processing*, pages 1–10.

Kakarala, R. (2012). Interpreting the phase spectrum in Fourier Analysis of partial ranking data. *Advances in Numerical Analysis*.

Kalai, G. (2002). A fourier-theoretic perspective on the condorcet paradox and arrow's theorem. *Advances in Applied Mathematics*, 29(3):412–426.

Kamishima, T. (2003). Nantonac collaborative filtering: recommendation based on order responses. In *KDD*, pages 583–588. ACM.

Kapicioglu, B., Rosenberg, D., Schapire, R. E., and Jebara, T. (2014). Collaborative ranking for local preferences. In *AISTATS*, pages 466–474.

Karpinski, M. and Schudy, W. (2010). Faster algorithms for feedback arc set tournament, kemeny rank aggregation and betweenness tournament. *Algorithms and Computation*, pages 3–14.

Keener, J. P. (1993). The perron-frobenius theorem and the ranking of football teams. *SIAM review*, 35(1):80–93.

Kemeny, J. (1959). Mathematics without numbers. *Daedalus*, 88:571–591.

Kendall, M. G. (1955). Further contributions to the theory of paired comparisons. *Biometrics*, 11(1):43–62.

Kendall, M. G. and Babington Smith, B. (1940). On the method of paired comparisons. *Biometrika*, 31(3/4):324–345.

Kenkre, S., Khan, A., and Pandit, V. (2011). On discovering bucket orders from preference data. In *Society for Industrial and Applied Mathematics. Proceedings of the SIAM International Conference on Data Mining*, page 872. SIAM.

Kenyon-Mathieu, C. and Schudy, W. (2007). How to rank with few errors. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 95–103. ACM.

Khetan, A. and Oh, S. (2016). Data-driven rank breaking for efficient rank aggregation. *arxiv preprint*.

Kidwell, P., Lebanon, G., and Cleveland, W. S. (2008). Visualizing incomplete and partially ranked data. *IEEE transactions on visualization and computer graphics*, 14(6):1356–63.

Kim, M., Farnoud, F., and Milenkovic, O. (2014). Hydra: gene prioritization via hybrid distance-score rank aggregation. *Bioinformatics*.

Knuth, D. E. (1973). *The Art of Computer Programming, Vol. 3: Sorting and Searching*.

Koczkodaj, W. and Orlowski, M. (1997). An orthogonal basis for computing a consistent approximation to a pairwise comparisons matrix. *Computers & Mathematics with Applications*, 34(10):41 – 47.

Koczkodaj, W. W. and Orłowski, M. (1999). Computing a consistent approximation to a generalized pairwise comparisons matrix. *Computers & Mathematics with Applications*, 37(3):79–85.

Kolde, R., Laur, S., Adler, P., and Vilo, J. (2012). Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*, 28(4):573–580.

Kondor, R. (2010). A fourier space algorithm for solving quadratic assignment problems. In *SODA*, pages 1017–1028. SIAM.

Kondor, R. and Barbosa, M. S. (2010). Ranking with kernels in Fourier space. In *Proceedings of COLT'10*, pages 451–463.

Kondor, R. and Dempsey, W. (2012). Multiresolution analysis on the symmetric group. In *Neural Information Processing Systems 25*.

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

Kondor, R., Teneva, N., and Garg, V. (2014). Multiresolution matrix factorization. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1620–1628.

Koppen, M. (1995). Random utility representation of binary choice probabilities: critical graphs yielding critical necessary conditions. *Journal of Mathematical Psychology*, 39(1):21–39.

Koren, Y. (2009). The bellkor solution to the netflix grand prize.

Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, (8):30–37.

Kuang, D., Shi, Z., Osher, S., and Bertozzi, A. (2016). A harmonic extension approach for collaborative ranking. *arXiv preprint*.

Lahaie, S. and Shah, N. (2014). Neutrality and geometry of mean voting. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 333–350. ACM.

Lawson, B. L., Orrison, M. E., and Uminsky, D. T. (2006). Spectral analysis of the supreme court. *Mathematics Magazine*, 79(5):340–346.

Lebanon, G. and Lafferty, J. (2002). Cranking: Combining rankings using conditional probability models on permutations. In *Proceedings of the 19th International Conference on Machine Learning*, pages 363–370.

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

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

Lee, J., Bengio, S., Kim, S., Lebanon, G., and Singer, Y. (2014). Local collaborative ranking. In *Proceedings of the 23rd international conference on World wide web*, pages 85–96. ACM.

Lee, P. H. and Yu, P. L. H. (2012). Mixtures of weighted distance-based models for ranking data with applications in political studies. *Computational Statistics & Data Analysis*, 56(8):2486 – 2500.

Lei, H., Xia, J., Guo, F., Zou, Y., Chen, W., and Liu, Z. (2016). Visual exploration of latent ranking evolutions in time series. *Journal of Visualization*, pages 1–13.

Lenk, P. J., DeSarbo, W. S., Green, P. E., and Young, M. R. (1996). Hierarchical bayes conjoint analysis: Recovery of partworth heterogeneity from reduced experimental designs. *Marketing Science*, 15(2):173–191.

Li, H. and Huh, W. T. (2011). Pricing multiple products with the multinomial logit and nested logit models: Concavity and implications. *Manufacturing & Service Operations Management*, 13(4):549–563.

Li, P., Wu, Q., and Burges, C. J. (2007). Mcrank: Learning to rank using multiple classification and gradient boosting. In *Advances in neural information processing systems*, pages 897–904.

Lim, C. H. and Wright, S. (2014). Beyond the birkhoff polytope: Convex relaxations for vector permutation problems. In *Advances in Neural Information Processing Systems*, pages 2168–2176.

Liqun, X. (2000). A multistage ranking model. *Psychometrika*, 65(2):217–231.

Liu, N. N. and Yang, Q. (2008). Eigenrank: a ranking-oriented approach to collaborative filtering. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 83–90. ACM.

Liu, T.-Y. (2009). Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331.

Lorena, L. H. N., Lorena, A. C., Lorena, L. A. N., and De Leon Carvalho, A. C. P. (2014). Clustering search applied to rank aggregation. In *Intelligent Systems (BRACIS), 2014 Brazilian Conference on*, pages 198–203. IEEE.

Lothaire, M. (1983). *Combinatorics on words*. Encyclopedia of mathematics and its applications. Addison-Wesley, Advanced Book Program, World Science Division.

Louviere, J. J. (1988). *Analyzing decision making: Metric conjoint analysis*. Number 67. Sage.

Louviere, J. J., Hensher, D. A., and Swait, J. D. (2000). *Stated choice methods: analysis and applications*. Cambridge University Press.

Lu, T. and Boutilier, C. (2011a). Learning mallows models with pairwise preferences. In *ICML*, pages 145–152.

Lu, T. and Boutilier, C. (2011b). Robust approximation and incremental elicitation in voting protocols. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 287.

Lu, T. and Boutilier, C. (2014). Effective sampling and learning for mallows models with pairwise-preference data. *Journal of Machine Learning Research*, 15:3783–3829.

Lu, Y. and Negahban, S. N. (2014). Individualized rank aggregation using nuclear norm regularization. *arXiv preprint*.

Luce, R. D. (1959). *Individual Choice Behavior*. Wiley.

Luce, R. D. (1977). The choice axiom after twenty years. *Journal of Mathematical Psychology*, 15(3):215 – 233.

Luce, R. D. and Suppes, P. (1965). *Preference, utility, and subjective probability*. Wiley.

Luce, R. D. and Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, 1(1):1 – 27.

Maire, M. (2010). Simultaneous segmentation and figure/ground organization using angular embedding. In *Computer Vision–ECCV 2010*, pages 450–464. Springer.

Mallat, S. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *Pattern Analysis and Machine Intelligence, IEEE*, II(7).

Mallat, S. (2008). *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*. Academic Press, 3rd edition.

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

Manski, C. F. (1977). The structure of random utility models. *Theory and Decision*, 8(3):229–254.

Mao, A., Procaccia, A. D., and Chen, Y. (2013). Better human computation through principled voting. In *AAAI*. Citeseer.

Marden, J. I. (1996). *Analyzing and Modeling Rank Data*. CRC Press, London.

Marley, A. A. J. (1968). Some probabilistic models of simple choice and ranking. *Journal of Mathematical Psychology*, 5(2):311–332.

Marlin, B. M., Zemel, R. S., Roweis, S., and Slaney, M. (2007). Collaborative filtering and the missing at random assumption. In *In Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence (UAI)*.

Maron, M. E. and Kuhns, J. L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the ACM (JACM)*, 7(3):216–244.

Marschak, J. (1959). Binary choice constraints on random utility indicators. Cowles Foundation Discussion Papers 74, Cowles Foundation for Research in Economics, Yale University.

Masarotto, G. and Varin, C. (2012). The ranking lasso and its application to sport tournaments. *The Annals of Applied Statistics*, 6(4):1949–1970.

Mattei, N. (2011). Empirical evaluation of voting rules with strictly ordered preference data. pages 165–177, Berlin, Heidelberg. Springer.

Maydeu-Olivares, A. (1999). Thurstonian modeling of ranking data via mean and covariance structure analysis. *Psychometrika*, pages 325–340.

Maystre, L. and Grossglauser, M. (2015a). Fast and accurate inference of plackett–luce models. In *Advances in Neural Information Processing Systems*, pages 172–180.

Maystre, L. and Grossglauser, M. (2015b). Robust active ranking from sparse noisy comparisons. *arxiv preprint*.

McCullagh, P. (1993). Models on spheres and models for permutations. In *Probability models and statistical analyses for ranking data*, pages 278–283. Springer.

McFadden, D. (1974a). Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, pages 105–142.

McFadden, D. (1974b). The measurement of urban travel demand. *Journal of public economics*, 3(4):303–328.

McFadden, D. (1980). Econometric models for probabilistic choice among products. *The Journal of Business*, 53(3):S13–S29.

McFadden, D. and Train, K. (2000). Mixed mnl models for discrete response. *Journal of applied Econometrics*, 15(5):447–470.

Meek, C. and Meila, M. (2014). Recursive inversion models for permutations. In *Advances in Neural Information Processing Systems 27*, pages 631–639.

Meila, M. and Bao, L. (2008). Estimation and clustering with infinite rankings. In *UAI*, pages 393–402.

Meila, M. and Chen, H. (2010). Dirichlet process mixtures of generalized mallows models. In *UAI 2010, Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA, July 8-11, 2010*, pages 358–367.

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

Meilă, M. and Bao, L. (2010). An exponential model for infinite rankings. *Journal of Macine Learning Research*, 11:3481–3518.

Mersmann, O., Preuss, M., Trautmann, H., Bischl, B., and Weihs, C. (2015). Analyzing the bbob results by means of benchmarking concepts. *Evol. Comput.*, 23(1):161–185.

Mollica, C. and Tardella, L. (2015). Bayesian mixture of plackett-luce models for partially ranked data. *arXiv preprint*.

Mosteller, F. (1951). Remarks on the method of paired comparisons: I. the least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, 16(1):3–9.

Murphy, T. B. and Martin, D. (2003). Mixtures of distance-based models for ranking data. *Computational statistics & data analysis*, 41(3):645–655.

Natarajan, K., Song, M., and Teo, C.-P. (2009). Persistency model and its applications in choice modeling. *Management Science*, 55(3):453–469.

Negahban, S., Oh, S., and Shah, D. (2012). Iterative ranking from pair-wise comparisons. In *Advances in Neural Information Processing Systems*, pages 2474–2482.

Netzer, O., Toubia, O., Bradlow, E., Dahan, E., Evgeniou, T., Feinberg, F., Feit, E., Hui, S., Johnson, J., Liechty, J., Orlin, J., and Rao, V. (2008). Beyond conjoint analysis: Advances in preference measurement. *Marketing Letters*, 19(3):337–354.

Nikolenko, S. and Sirotkin, A. (2011). A new bayesian rating system for team competitions. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 601–608.

Niu, S., Lan, Y., Guo, J., and Cheng, X. (2013). Stochastic rank aggregation. In *Proceedings of the Twenty-Ninth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-13)*, pages 478–487. AUAI Press.

Nunnally, J. C., Bernstein, I. H., and Berge, J. M. F. (1967). *Psychometric theory*, volume 226. JSTOR.

Oh, S. and Shah, D. (2014). Learning mixed multinomial logit model from ordinal data. In *Advances in Neural Information Processing Systems*, pages 595–603.

Oh, S., Thekumparampil, K. K., and Xu, J. (2015). Collaboratively learning preferences from ordinal data. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 1909–1917. Curran Associates, Inc.

Osting, B., Brune, C., and Osher, S. (2013). Enhanced statistical rankings via targeted data collection. In *Journal of Machine Learning Research, W&CP (ICML 2013)*, volume 28 (1), pages 489–497.

Pachauri, D., Collins, M., Kondor, R., and Singh, V. (2012). Incorporating domain knowledge in matching problems via harmonic analysis. In *ICML 2012*.

Pachauri, D., Kondor, R., Sargur, G., and Singh, V. (2014). Permutation diffusion maps (pdm) with application to the image association problem in computer vision. In *Advances in Neural Information Processing Systems*, pages 541–549.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Technical report.

Pahikkala, T., Tsivtsivadze, E., Airola, A., Boberg, J., and Salakoski, T. (2007). Learning to rank with pairwise regularized least-squares. In *SIGIR 2007 workshop on learning to rank for information retrieval*, volume 80, pages 27–33. Citeseer.

Pardalos, P. M., Rendl, F., and Wolkowicz, H. (1994). *Quadratic Assignment and Related Problems: DIMACS Workshop, May 20-21, 1993*, volume 16. American Mathematical Soc.

Pareek, H. H. and Ravikumar, P. K. (2014). A representation theory for ranking functions. In *Advances in Neural Information Processing Systems*, pages 361–369.

Park, D., Neeman, J., Zhang, J., Sanghavi, S., and Dhillon, I. (2015). Preference completion: Large-scale collaborative ranking from pairwise comparisons. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1907–1916.

Patel, T., Telesca, D., Rallo, R., George, S., Xia, T., and Nel, A. E. (2013). Hierarchical rank aggregation with applications to nanotoxicology. *Journal of Agricultural, Biological, and Environmental Statistics*, 18(2):159–177.

Patil, G. P. and Taillie, C. (2004). Multiple indicators, partially ordered sets, and linear extensions: Multi-criterion ranking and prioritization. *Environmental and Ecological Statistics*, 11(2):199–228.

Pfeiffer, T., Gao, X. A., Chen, Y., Mao, A., and Rand, D. G. (2012). Adaptive polling for information aggregation. In *AAAI*.

Plackett, R. L. (1975). The analysis of permutations. *Applied Statistics*, 2(24):193–202.

Plis, S. M., Mccracken, S., Lane, T., and Calhoun, V. D. (2011). Directional statistics on permutations. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS-11)*, volume 15, pages 600–608. Journal of Machine Learning Research - Workshop and Conference Proceedings.

Popov, S., Popova, A., and Regenwetter, M. (2014). Consensus in organizations: Hunting for the social choice conundrum in apa elections. *Decision*, 1(2):123–146.

Prasad, A., Pareek, H., and Ravikumar, P. (2015). Distributional rank aggregation, and an axiomatic analysis. In Blei, D. and Bach, F., editors, *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2104–2112. JMLR Workshop and Conference Proceedings.

Prati, R. C. (2012). Combining feature ranking algorithms through rank aggregation. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1–8. IEEE.

Procaccia, A. D., Reddi, S., and Shah, N. (2012). A maximum likelihood approach for selecting sets of alternatives. In *Proceedings of the Twenty-Eighth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-12)*, pages 695–704, Corvallis, Oregon. AUAI Press.

Procaccia, A. D. and Rosenschein, J. S. (2006). Junta distributions and the average-case complexity of manipulating elections. In *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*, pages 497–504. ACM.

Procaccia, A. D. and Shah, N. (2015). Is approval voting optimal given approval votes? In *Advances in Neural Information Processing Systems*, pages 1792–1800.

Pujari, M. and Kanawati, R. (2012). Supervised rank aggregation approach for link prediction in complex networks. In *Proceedings of the 21st international conference companion on world wide web*, pages 1189–1196. ACM.

Qin, T., Geng, X., and Liu, T.-Y. (2010). A new probabilistic model for rank aggregation. In *Advances in Neural Information Processing Systems 23*, pages 1948–1956.

Ragnarsson, K. and Tenner, B. E. (2011). Homology of the boolean complex. *Journal of Algebraic Combinatorics*, 34(4):617–639.

Rajkumar, A. and Agarwal, S. (2014). A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In *Proceedings of the 31st International Conference on Machine Learning*.

Rajkumar, A., Ghoshal, S., Lim, L.-H., and Agarwal, S. (2015). Ranking from stochastic pairwise preferences: Recovering condorcet winners and tournament solution sets at the top. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 665–673.

Raman, K. and Joachims, T. (2015). Bayesian ordinal peer grading. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*, pages 149–156. ACM.

Rao, P. and Kupper, L. (1967). Ties in paired-comparison experiments: A generalization of the bradley-terry model. *Journal of the American Statistical Association*, 62(317):194–204.

Ravikumar, P. D., Tewari, A., and Yang, E. (2011). On ndcg consistency of listwise ranking methods. In *International conference on artificial intelligence and statistics*, pages 618–626.

Regenwetter, M. and Marley, A. A. J. (2001). Random relations, random utilities, and random functions. *Journal of Mathematical Psychology*, 45(6):864–912.

Reinelt, G. (1985). *The linear ordering problem: algorithms and applications*, volume 8. Heldermann.

Reiner, V., Saliola, F., and Welker, V. (2014). Spectra of symmetrized shuffling operators. *Memoirs of the American Mathematical Society*, 228(1072).

Reiner, V. and Webb, P. (2004). Combinatorics of the bar resolution in group cohomology. *J. Pure Appl. Algebra*, 190:291–327.

Renda, M. E. and Straccia, U. (2003). Web metasearch: rank vs. score based rank aggregation methods. In *Proceedings of the 2003 ACM symposium on Applied computing*, pages 841–846. ACM.

Rendle, S., Freudenthaler, C., Gantner, Z., and Schmidt-Thieme, L. (2009). Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 452–461. AUAI Press.

Renteln, P. (2011). The distance spectra of cayley graphs of coxeter groups. *Discrete Mathematics*, 311(8–9):738 – 755.

Restle, F. (1961). Psychology of judgment and choice: A theoretical essay.

Reutenauer, C. (1993). *Free Lie algebras, volume 7 of London Mathematical Society Monographs. New Series*. The Clarendon Press Oxford University Press, New York.

Risse, M. (2005). Why the count de borda cannot beat the marquis de condorcet. *Social Choice and Welfare*, 25(1):95–113.

Roy, B. (1968). Classement et choix en présence de points de vue multiples. *Revue française d'automatique, d'informatique et de recherche opérationnelle. Recherche opérationnelle*, 2(1):57–75.

Roy, B. (1991). The outranking approach and the foundations of electre methods. *Theory and decision*, 31(1):49–73.

Rustamov, R. M. and Guibas, L. J. (2013). Wavelets on graphs via deep learning. In *Advances in Neural Information Processing Systems 26.*, pages 998–1006.

Ryan, M. (1999). Using conjoint analysis to take account of patient preferences and go beyond health outcomes: an application to in vitro fertilisation. *Social Science & Medicine*, 48(4):535 – 546.

Ryan, M. and Farrar, S. (2000). Using conjoint analysis to elicit preferences for health care. *British Medical Journal*, 320(7248):1530.

Saari, D. G. (2000). Mathematical structure of voting paradoxes. *Economic Theory*, 15(1):1–53.

Saari, D. G. (2005). The profile structure for luce's choice axiom. *Journal of Mathematical Psychology*, 49(3):226–253.

Saari, D. G. and Merlin, V. R. (2000). A geometric examination of kemeny's rule. *Social Choice and Welfare*, 17(3):403–438.

Saaty, T. L. (1977). A scaling method for priorities in hierarchical structures. *Journal of mathematical psychology*, 15(3):234–281.

Sagan, B. (2013). *The symmetric group: representations, combinatorial algorithms, and symmetric functions*, volume 203. Springer Science & Business Media.

Salakhutdinov, R., Mnih, A., and Hinton, G. (2007). Restricted boltzmann machines for collaborative filtering. In *Proceedings of the 24th international conference on Machine learning*, pages 791–798. ACM.

Samuelson, L. (1985). On the independence from irrelevant alternatives in probabilistic choice models. *Journal of Economic Theory*, 35(2):376–389.

Satterthwaite, M. A. (1975). Strategy-proofness and arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of economic theory*, 10(2):187–217.

Schalekamp, F. and van Zuylen, A. (2009). Rank aggregation: Together we're strong. In *Proceedings of the Eleventh Workshop on Algorithm Engineering and Experiments*, pages 38–51.

Schultz, M. and Joachims, T. (2004). Learning a distance metric from relative comparisons. In Thrun, S., Saul, L. K., and Schölkopf, B., editors, *Advances in Neural Information Processing Systems 16*, pages 41–48. MIT Press.

Sen, A. (1970). Collective choice and social welfare.

Sen, A. (1977). On weights and measures: informational constraints in social welfare analysis. *Econometrica: Journal of the Econometric Society*, pages 1539–1572.

Shadi, K., Bakhshi, S., Gutman, D. A., Mayberg, H. S., and Dovrolis, C. (2015). A symmetry-based method to infer structural brain networks from tractography data. *arXiv preprint*.

Shah, N. B., Balakrishnan, S., Guntuboyina, A., and Wainright, M. J. (2015a). Stochastically transitive models for pairwise comparisons: Statistical and computational issues. *arXiv preprint*.

Shah, N. B., Parekh, A., Balakrishnan, S., Ramchandran, K., Bradley, J., and Wainwright, M. (2015b). Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 856–865.

Shah, N. B. and Wainwright, M. J. (2015). Simple, robust and optimal ranking from pairwise comparisons. *arXiv preprint*.

Shalev-Shwartz, S. and Singer, Y. (2007). A unified algorithmic approach for efficient online label ranking. In *AISTATS*, pages 452–459.

Shani, G. and Gunawardana, A. (2011). Evaluating recommendation systems. In *Recommender systems handbook*, pages 257–297. Springer.

Shashua, A. and Levin, A. (2002). Ranking with large margin principle: Two approaches. In *Advances in neural information processing systems*, pages 937–944.

Shi, C., Cui, W., Liu, S., Xu, P., Chen, W., and Qu, H. (2012). Rankexplorer: Visualization of ranking changes in large time series data. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2669–2678.

Sibony, E. (2014). Borda count approximation of Kemeny's rule and pairwise voting inconsistencies. In *Proceedings of the NIPS 2014 Workshop on Analysis of Rank Data*.

Sibony, E., Clemencon, S., and Jakubowicz, J. (2014). Multiresolution analysis of incomplete rankings with applications to prediction. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 88–95.

Sibony, E., Clémençon, S., and Jakubowicz, J. (2015). Mra-based statistical learning from incomplete rankings. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 1432–1441.

Sibony, E., Clémençon, S., and Jakubowicz, J. (2016). A multiresolution analysis framework for the statistical analysis of incomplete rankings. *arXiv preprint*.

Slater, P. (1961). Inconsistencies in a schedule of paired comparisons. *Biometrika*, 48(3/4):303–312.

Soufiani, H., Parkes, D., and Xia, L. (2013a). Preference elicitation for general random utility models. In *Proceedings of the Twenty-Ninth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-13)*, pages 596–605, Corvallis, Oregon. AUAI Press.

Soufiani, H. A., Chen, W., Parkes, D. C., and Xia, L. (2013b). Generalized method-of-moments for rank aggregation. In *Advances in Neural Information Processing Systems 26*, pages 2706–2714.

Soufiani, H. A., Parkes, D. C., and Xia, L. (2014a). Computing parametric ranking models via rank-breaking. In *Proceedings of The 31st International Conference on Machine Learning*. International Conference on Machine Learning.

Soufiani, H. A., Parkes, D. C., and Xia, L. (2014b). A statistical decision-theoretic framework for social choice. In *Advances in Neural Information Processing Systems*, pages 3185–3193.

Soutar, G. N. and Turner, J. P. (2002). Students' preferences for university: a conjoint analysis. *International Journal of Educational Management*, 16(1):40–45.

Stanley, R. P. (1986). *Enumerative Combinatorics*. Wadsworth Publ. Co., Belmont, CA, USA.

Stoyanovich, J., Jacob, M., and Gong, X. (2015). Analyzing crowd rankings. In *Proceedings of the 18th International Workshop on Web and Databases*, pages 41–47. ACM.

Sturmfels, B. and Welker, V. (2012). Commutative algebra of statistical ranking. *Journal of Algebra*, 361:264 – 286.

Suck, R. (1992). Geometric and combinatorial properties of the polytope of binary choice probabilities. *Mathematical Social Sciences*, 23(1):81–102.

Sun, M., Lebanon, G., and Collins-Thompson, K. (2010). Visualizing differences in web search algorithms using the expected weighted hoeffding distance. In *Proceedings of the 19th international conference on World wide web*, pages 931–940. ACM.

Sun, M., Lebanon, G., and Kidwell, P. (2012). Estimating probabilities in recommendation systems. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(3):471–492.

Swait, J. and Louviere, J. (1993). The role of the scale parameter in the estimation and comparison of multinomial logit models. *Journal of Marketing Research*, pages 305–314.

Tabourier, L., Libert, A.-S., and Lambiotte, R. (2014). Rankmerging: Learning to rank in large-scale social networks. In *DyNakII, 2nd International Workshop on Dynamic Networks and Knowledge Discovery (PKDD 2014 workshop)*.

Takahashi, R. and Morimura, T. (2015). Predicting preference reversals via gaussian process uncertainty aversion. In *AISTATS*.

Takane, Y. (1987). Analysis of covariance structures and probabilistic binary choice data. *Communication & Cognition*, 20:45–62.

Talluri, K. and Van Ryzin, G. (2004). Revenue management under a general discrete choice model of consumer behavior. *Management Science*, 50(1):15–33.

Tan, A. C., Naiman, D. Q., Xu, L., Winslow, R. L., and Geman, D. (2005). Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*, 21(20):3896–3904.

Tax, N., Bockting, S., and Hiemstra, D. (2015). A cross-benchmark comparison of 87 learning to rank methods. *Information Processing & Management*, 51(6):757–772.

Taylor, M., Guiver, J., Robertson, S., and Minka, T. (2008). Softrank: optimizing non-smooth rank metrics. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 77–86. ACM.

Thurstone, L. L. (1927a). A law of comparative judgment. *Psychological Review*, 34(4):273–286.

Thurstone, L. L. (1927b). The method of paired comparisons for social values. *The Journal of Abnormal and Social Psychology*, 21(4):384.

Tideman, N. (2006). *Collective decisions and voting: the potential for public choice*. Ashgate Publishing, Ltd.

Titchener, E. B. (1901). *Experimental Psychology: A Manual of Laboratory Practice*. Number 1 in Experimental Psychology. MacMillan.

Torgerson, W. S. (1958). Theory and methods of scaling.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press.

Tran, T. and Venkatesh, S. (2014). Permutation models for collaborative ranking. *arXiv preprint arXiv:1407.6128*.

Tsai, R.-C. and Yao, G. (2000). Testing thurstonian case v ranking models using posterior predictive checks. *British Journal of Mathematical and Statistical Psychology*, 53(2):275–292.

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological review*, 79(4):281.

Tversky, A. and Russo, J. E. (1969). Substitutability and similarity in binary choices. *Journal of Mathematical Psychology*, 6(1):1–12.

Ukkonen, A. (2007). *Advances in Intelligent Data Analysis VII: 7th International Symposium on Intelligent Data Analysis, IDA 2007, Ljubljana, Slovenia, September 6-8, 2007. Proceedings*, chapter Visualizing Sets of Partial Rankings, pages 240–251. Springer Berlin Heidelberg, Berlin, Heidelberg.

Uyemura-Reyes, J.-C. (2002). *Random walks, semidirect products, and card shuffling*. PhD thesis, Stanford University.

Van Zuylen, A. and Williamson, D. P. (2007). Deterministic algorithms for rank aggregation and other ranking and clustering problems. In *Approximation and Online Algorithms*, pages 260–273. Springer.

Van Zuylen, A. and Williamson, D. P. (2009). Deterministic pivoting algorithms for constrained ranking and clustering problems. *Mathematics of Operations Research*, 34(3):594–620.

Velasquez, M. and Hester, P. T. (2013). An analysis of multi-criteria decision making methods. *International Journal of Operations Research*, 10(2):56–66.

Viana, M. (2006). Symmetry studies and decompositions of entropy. *Entropy*, 8(2):88–109.

Volkovs, M. and Zemel, R. S. (2012). Collaborative ranking with 17 parameters. In *Advances in Neural Information Processing Systems*, pages 2294–2302.

Volkovs, M. N. and Zemel, R. S. (2014). New learning methods for supervised and unsupervised preference aggregation. *Journal of Machine Learning Research*, 15:1135–1176.

Wagstaff, K., Cardie, C., Rogers, S., and Schrödl, S. (2001). Constrained k-means clustering with background knowledge. In *ICML*, volume 1, pages 577–584.

Walker, J. and Ben-Akiva, M. (2002). Generalized random utility model. *Mathematical Social Sciences*, 43(3):303–343.

Wallenius, J., Dyer, J. S., Fishburn, P. C., Steuer, R. E., Zionts, S., and Deb, K. (2008). Multiple criteria decision making, multiattribute utility theory: recent accomplishments and what lies ahead. *Management science*, 54(7):1336–1349.

Walsh, T. (2014). The peerrank method for peer assessment. In *ECAI 2014 - 21st European Conference on Artificial Intelligence, 18-22 August 2014, Prague, Czech Republic - Including Prestigious Applications of Intelligent Systems (PAIS 2014)*, pages 909–914.

Wang, D., Mazumdar, A., and Wornell, G. W. (2013). A rate-distortion theory for permutation spaces. In *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, pages 2562–2566. IEEE.

Wang, J., Srebro, N., and Evans, J. (2014). Active collaborative permutation learning. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 502–511. ACM.

Wauthier, F., Jordan, M., and Jojic, N. (2013). Efficient ranking from pairwise comparisons. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 109–117.

Webber, W., Moffat, A., and Zobel, J. (2010). A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):20.

Weimer, M., Karatzoglou, A., Le, Q. V., and Smola, A. (2007). Maximum margin matrix factorization for collaborative ranking. *Advances in neural information processing systems*, pages 1–8.

Weng, R. C. and Lin, C.-J. (2011). A bayesian approximation method for online ranking. *Journal of Machine Learning Research*, 12:267–300.

Williams, H. C. W. L. (1977). On the formation of travel demand models and economic evaluation measures of user benefit. *Environment and planning A*, 9(3):285–344.

Wilson, A., Fern, A., and Tadepalli, P. (2012). A bayesian approach for policy learning from trajectory preference queries. In *Advances in neural information processing systems*, pages 1133–1141.

Wittink, D. R. and Cattin, P. (1989). Commercial use of conjoint analysis: An update. *Journal of Marketing*, 53(3):91–96.

Wong, H. S., Chin, T.-J., Yu, J., and Suter, D. (2013). Mode seeking over permutations for rapid geometric model fitting. *Pattern Recognition*, 46(1):257–271.

Wu, O., Zuo, H., Zhu, M., Hu, W., Gao, J., and Wang, H. (2009). Rank aggregation based text feature selection. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '09, pages 165–172, Washington, DC, USA. IEEE Computer Society.

Wu, R., Xu, J., Srikant, R., Massoulié, L., Lelarge, M., and Hajek, B. (2015). Clustering and inference from pairwise comparisons. In *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pages 449–450. ACM.

Xia, F., Liu, T.-Y., Wang, J., Zhang, W., and Li, H. (2008). Listwise approach to learning to rank: theory and algorithm. In *Proceedings of the 25th international conference on Machine learning*, pages 1192–1199. ACM.

Xia, L. and Conitzer, V. (2008). Generalized scoring rules and the frequency of coalitional manipulability. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 109–118. ACM.

Xu, B., Bu, J., Chen, C., Cai, D., He, X., Liu, W., and Luo, J. (2011a). Efficient manifold ranking for image retrieval. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 525–534. ACM.

Xu, J. and Li, H. (2007). Adarank: a boosting algorithm for information retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 391–398. ACM.

Xu, Q., Huang, Q., Jiang, T., Yan, B., Lin, W., and Yao, Y. (2012). Hodgerank on random graphs for subjective video quality assessment. *Multimedia, IEEE Transactions on*, 14(3):844–857.

Xu, Q., Jiang, T., Yao, Y., Huang, Q., Yan, B., and Lin, W. (2011b). Random partial paired comparison for subjective video quality assessment via hodgerank. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 393–402. ACM.

Xu, Q., Xiong, J., Huang, Q., and Yao, Y. (2014). Robust statistical ranking: Theory and algorithms. *arXiv preprint*.

Yao, G. and Böckenholt, U. (1999). Bayesian estimation of thurstonian ranking models based on the gibbs sampler. *British Journal of Mathematical and Statistical Psychology*, 52(1):79–92.

Yasutake, S., Hatano, K., Takimoto, E., and Takeda, M. (2012). Online rank aggregation. In *ACML*, pages 539–553.

Yellott, J. I. (1977). The relationship between luce's choice axiom, thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, 15(2):109–144.

Yi, J., Jin, R., Jain, S., and Jain, A. (2013). Inferring users' preferences from crowdsourced pairwise comparisons: A matrix completion approach. In *First AAAI Conference on Human Computation and Crowdsourcing*.

Young, H. (1988). Condorcet's theory of voting. *American Political Science Review*, 82(4):1231–1244.

Young, H. P. (1977). Extending condorcet's rule. *Journal of Economic Theory*, 16(2):335–353.

Young, H. P. and Levenglick, A. (1978). A consistent extension of condorcet's election principle. *SIAM Journal on applied Mathematics*, 35(2):285–300.

Yu, P. L. H. and Chan, L. K. Y. (2001). Bayesian analysis of wandering vector models for displaying ranking data. *Statistica Sinica*, 11(2):445–461.

Yu, P. L. H., Lam, K. F., and Alvo, M. (2002). Nonparametric rank test for independence in opinion surveys. *Australian Journal of Statistics*, 31:279–290.

Yu, S. X. (2009). Angular embedding: from jarring intensity differences to perceived luminance. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2302–2309. IEEE.

Yu, S. X. (2012). Angular embedding: A robust quadratic criterion. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(1):158–173.

Yue, Y., Finley, T., Radlinski, F., and Joachims, T. (2007). A support vector method for optimizing average precision. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 271–278. ACM.

Zeleny, M. and Cochrane, J. L. (1973). *Multiple criteria decision making*. University of South Carolina Press.

Zermelo, E. (1929). Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 29(1):436–460.

Zhang, D. and Cooper, W. L. (2005). Revenue management for parallel flights with customer-choice behavior. *Operations Research*, 53(3):415–431.

Zhang, J. (2004). Binary choice, subset choice, random utility, and ranking: A unified perspective using the permutahedron. *Journal of Mathematical Psychology*, 48(2):107–134.

Zheng, W.-S., Gong, S., and Xiang, T. (2013). Reidentification by relative distance comparison. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(3):653–668.

Zhou, D., Weston, J., Gretton, A., Bousquet, O., and Schölkopf, B. (2004). Ranking on data manifolds. *Advances in neural information processing systems*, 16:169–176.

Zwicker, W. S. (2008). Consistency without neutrality in voting rules: When is a vote an average? *Mathematical and Computer Modelling*, 48(9):1357–1373.

# Analyse multirésolution de données de classements

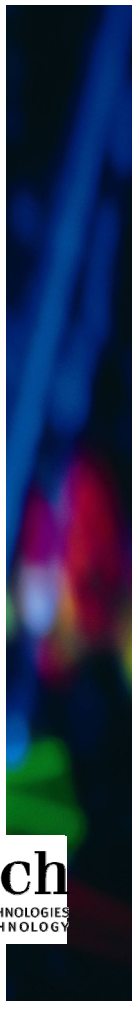
**Éric SIBONY**

**RESUME :** Cette thèse introduit un cadre d'analyse multirésolution pour les données de classements. Initiée au $18^e$ siècle dans le contexte d'élections, l'analyse des données de classements a attiré un intérêt majeur dans de nombreux domaines de la littérature scientifique : psychométrie, statistiques, économie, recherche opérationnelle, apprentissage automatique ou choix social computationel entre autres. Elle a de plus été revitalisée par des applications modernes comme les systèmes de recommandation, où le but est d'inférer les préférences des utilisateurs pour leur proposer les meilleures suggestions personnalisées. Dans ces contextes, les utilisateurs expriment leurs préférences seulement sur des petits sous-ensembles d'objets variant au sein d'un large catalogue. L'analyse de tels *classements incomplets* pose cependant un défi important, tant du point de vue statistique que computationnel, poussant les acteurs industriels à utiliser des méthodes qui n'exploitent qu'une partie de l'information disponible. Cette thèse introduit une nouvelle représentation pour les données, qui surmonte par construction ce double défi. Bien qu'elle repose sur des résultats de combinatoire et de topologie algébrique, ses nombreuses analogies avec l'analyse multirésolution en font un cadre naturel et efficace pour l'analyse des classements incomplets. Ne faisant aucune hypothèse sur les données, elle mène déjà à des estimateurs au-delà de l'état-de-l'art pour des petits catalogues d'objets et peut être combinée avec de nombreuses procédures de régularisation pour des larges catalogues. Pour toutes ces raisons, nous croyons que cette représentation multirésolution ouvre la voie à de nombreux développements et applications futurs.

**MOTS-CLEFS :** Classements, Apprentissage des préférences, Analyse multirésolution, Ondelettes

**ABSTRACT :** This thesis introduces a multiresolution analysis framework for ranking data. Initiated in the $18^{th}$ century in the context of elections, the analysis of ranking data has attracted a major interest in many fields of the scientific literature : psychometry, statistics, economics, operations research, machine learning or computational social choice among others. It has been even more revitalized by modern applications such as recommender systems, where the goal is to infer users preferences in order to make them the best personalized suggestions. In these settings, users express their preferences only on small and varying subsets of a large catalog of items. The analysis of such *incomplete rankings* poses however both a great statistical and computational challenge, leading industrial actors to use methods that only exploit a fraction of available information. This thesis introduces a new representation for the data, which by construction overcomes the two aforementioned challenges. Though it relies on results from combinatorics and algebraic topology, it shares several analogies with multiresolution analysis, offering a natural and efficient framework for the analysis of incomplete rankings. As it does not involve any assumption on the data, it already leads to over-performing estimators in small-scale settings and can be combined with many regularization procedures for large-scale settings. For all those reasons, we believe that this multiresolution representation paves the way for a wide range of future developments and applications.

**KEY-WORDS :** Rankings, Preference learning, Multiresolution analysis, Wavelets



TELECOM ParisTech

EDiTE

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY