Grenoble Institute of Technology - Ensimag

Master thesis

Ranking Aggregation for Portfolio Selection

Léo Nicoletti

Master of Science in Industrial and Applied Mathematics (2015/2016)

1st March, 2016 - 26th August, 2016

SAP SE
Dietmar-Hopp-Allee 16
69190 Walldorf Germany

Internship supervisor
Selim Gökay
School supervisor
Olivier François

Abstract

This master thesis aims to design a general learning framework for financial mathematics by adapting Learning to Rank algorithms to learn from aggregated time series. For a given financial problem, the task is to rank strategies with respect to various performances measures to offer an optimal decision. Most often, the wide range of possible strategies offers an immense choice of alternatives and the number of total rankings scales factorially in the number of alternatives to be ranked. It motivates the use of advanced learning techniques to rank them in reasonable time. We propose to aggregate information coming from a wide variety of performance measures to independently score alternatives and provide supervision to learning algorithms. The work of J.C. Duchi *et al.*, *The Asymptotics of Ranking Algorithms* [1] lists assumptions needed to perform optimization of a learning model with consistent surrogates losses including aggregation of measures. A considerable amount of work is devoted to provide robust solutions of rank aggregation and Machine Learning, through testing, comparing and benchmarking algorithms. An application to the problem of bond portfolio selection for fitting the yield curve is provided at the end of the thesis.

Keywords – Rank aggregation, Learning to Rank, portfolio selection, Cross Entropy Monte Carlo method, Kemeny distance, Random Utility theory, Condorcet model, Mallows model, Plackett-Luce model, Normalized Discounted Cumulative Gain.

Contents

1 Formalization of the rank aggregation problem

1.1 Notations

We are provided with a finite set of n strategies to be ranked, by labeling them with naturals to use them as indices, we work with the set $[\![1, n]\!] := \{1, \cdots, n\}$. In order theory, preferences over a set are formalized with binary relations. We denote $\mathcal{R}$ the set of *complete weak orders* or *rankings* (reflexive, transitive binary relations) on $[\![1, n]\!]$. Rankings orders are not total since two alternatives can have the same ranks, rankings can have ties *eg.* $1 \succ 2, 3 \succ 4$.

The *total orders* on $[\![1, n]\!]$ (reflexive, transitive, antisymmetric, total binary relations) can be represented as permutations $\pi \in \mathfrak{S}_n$ in the symmetric group. For a given total order π, the set of alternatives $([\![1, n]\!], \succ_\pi)$ is a totally ordered set where the infix $\succ_\pi$ denotes the corresponding binary relation. For both partial and total orders, we use the permutation notation for convenience where r_i and π_i represent the rank of alternative i in the partial order r and in the total order π respectively. Therefore, the conditions $i \succ_r j$ and $i \succ_\pi j$ respectively write as $r_i < r_j$ and $\pi_i < \pi_j$.

We consider preferences provided by L different performance measures on the strategies, *ie.* L voters contributing with L total rankings. This framework is called *listwise* since each voter $l \in [\![1, L]\!]$ is able to judge each of the alternatives. Such listwise preferences form a *profile of total rankings* $\mathcal{D}_L = \{\pi^{(1)}, \cdots, \pi^{(L)}\}$. In a more general framework we could consider partial rankings or *pairwise* judgment preferences, as commonly used in quantitative psychology. Given a profile of rankings $\mathcal{D}_L$, a *rank aggregation scheme* - also called a *social welfare function* $F : \mathfrak{S}_n^L \to \mathfrak{S}_n$ aggregates their choices into a single desired total ranking. The social welfare function corresponds to an aggregation scheme in the world of rankings.

1.2 Distributional hypothesis

The natural assumption consists into considering observed data from a probability model. The hypothesis can be either assumed on the scores - *cardinal distribution* - or directly on the permutation space - *ordinal distribution*. Following the work of Prasad *et al.* (2015, [2]), we consider the latter hypothesis to specify desirable axioms of any aggregation scheme.

Distributional assumption. Let's denote $\mathcal{P}_n$ the set of all distributions over $\mathfrak{S}_n$. For any profile of rankings $\mathcal{D}_L = \{\pi^{(1)}, \cdots, \pi^{(L)}\}$, there exists $p \in \mathcal{P}_n$ such that

$$\forall l \in \{1, \cdots, L\},\ \pi^{(l)} \sim^{iid} p.$$

This assumption naturally satisfies the principle of *anonymity*, *ie.* the alternatives can be labelled differently without changing the known information since the aggregation scheme is fed with histogram information only. It is coined *distributional rank aggregation* by Prasad *et al.* and gives a convenient formalism for computing marginal probabilities provided with an appropriated model for p. A wide range of models are proposed around the concept of marginals, *eg.* Mallows φ-models, Condorcet models, Plackett-Luce model, that will be studied later.

The marginal probability of alternative i being ranked first is $\sum_{\pi \mid \pi_i = 1} p(\pi)$. The marginal probability of alternative i being ranked better than j is $p_{i \succ j} := \sum_{\pi \mid \pi_i < \pi_j} p(\pi)$. The asymptotic equivalent of the rank aggregation scheme is a map $F^* : \mathcal{P}_n \to \mathfrak{S}_n$ called a *distributional rank aggregation scheme* which outputs a consensus ranking $F^*(p) \in \mathfrak{S}_n$. The probability distribution p is called a *ranking model* since its knowledge suffices to define a rank aggregation scheme, by selecting a mode of the model

$$F^*(p) \in \underset{\pi \in \mathfrak{S}_n}{\arg\max}\ p(\pi).$$

1.3 Condorcet Criterion

In a publication that laid the groundwork of future research in the field of rankings, Condorcet (1785, [3]) proposed a desirable criterion for the aggregated ranking. $F(\mathcal{D}_L)_i$ naturally refers to the rank of alternative i in the order produced by the aggregation scheme F. For a given $\mathcal{D}_L$ and aggregation scheme F, the aggregation relation $\succ_{F(\mathcal{D}_L)}$ is defined by $\forall i, j \in \llbracket 1, n\rrbracket, i \succ_{F(\mathcal{D}_L)} j \iff F(\mathcal{D}_L)_i < F(\mathcal{D}_L)_j$.

Assuming that each voter l provides a set of pairwise preferences, we define the *majority relation* on the alternatives $\llbracket 1, n\rrbracket$ with the marginal weights

$$v_{i,j}(\mathcal{D}_L) := |\{l \in \llbracket 1, L\rrbracket \mid \pi_i^{(l)} < \pi_j^{(l)}\}|.$$

We define $i \succeq_M j \iff v_{i,j}(\mathcal{D}_L) \geq v_{j,i}(\mathcal{D}_L)$, $i \succ_M j \iff v_{i,j}(\mathcal{D}_L) > v_{j,i}(\mathcal{D}_L)$ and the *ex æquo relation* $i =_M j \iff v_{i,j}(\mathcal{D}_L) = v_{j,i}(\mathcal{D}_L)$. If the relation $\succ_M$ yields a total order, it is called the *Condorcet ranking* but in most of the cases the majority relation is not transitive, which compels the use of more sophisticated aggregation schemes. Note that the majority relation can be generalized to give more or less relevance to a voter by adding weights on pairwise preferences.

To comply with the usual graph approach in ranking, we explicit the direct graph associated with the majority relation as it is used in some algorithms. We first define the marginal sets $\mathcal{D}_{i\succ j} := \{\pi^{(l)} \mid l \in \llbracket 1, L\rrbracket, i \succ_{\pi^{(l)}} j\}$. The graph corresponding is often referred as the *weighted majority graph* in the literature and is formally defined as the directed graph $WMG(\mathcal{D}_L) = (\llbracket 1, n\rrbracket, \mathcal{W})$ where

$$\mathcal{W} = \{w_{i\to j} := |\mathcal{D}_{i\succ j}| \mid i \neq j, |\mathcal{D}_{i\succ j}| > 0\}.$$

The Condorcet Criterion (CC) ensures that a *Condorcet winner*, *ie.* an alternative i ranked first by a majority of voters, ends up first in the aggregated total ranking

$$\forall i, j \in \llbracket 1, n\rrbracket, i \succ_M j \implies (F(\mathcal{D}_L)_i = 1) \text{ and } (F(\mathcal{D}_L)_j > 1). \tag{CC}$$

This can be simply written as $\forall i, j \in \llbracket 1, n\rrbracket, i \succ_M j \implies i \succ_{F(\mathcal{D}_L)} j$. This criterion is weakened to deal with cycles in the majority relation $\succeq_M$ using partitions forming coherent *bins* of alternatives. We define the set $\mathcal{P}_0(\llbracket 1, n\rrbracket)$ of partitions $\{X_1, \cdots, X_r\}$, $\llbracket 1, n\rrbracket = \bigsqcup_{k=1}^{r} X_k$ such as $\forall k < l, \forall i \in X_k, \forall j \in X_l, i \succeq_M j$. The Extended Condorcet Criterion (XCC) is that for any partition $X \in \mathcal{P}_0(\llbracket 1, n\rrbracket)$ we have,

$$\forall k < l, \forall i \in X_k, \forall j \in X_l, i \succ_{F(\mathcal{D}_L)} j. \tag{XCC}$$

We say that the partition $X' = \{X'_1, \cdots, X'_{r'}\}$, $\llbracket 1, n\rrbracket = \bigsqcup_{l=1}^{r'} X'_l$ is *finer* than X if

$$\forall l \in \llbracket 1, r'\rrbracket, \exists k \in \llbracket 1, r\rrbracket, X'_l \subseteq X_k.$$

1.4 Social choice axioms

Arrow (1951, [4]) laid the axiomatic foundations of the social choice theory by proposing the five desired axioms which should be satisfied by any rank aggregation scheme F.

(S1) Non-dictatorship. The rank aggregation scheme must account for the preferences of multiple voters. It cannot simply mimic the preferences of a single voter.

(S2) Universality. For any set of rankings $\mathcal{D}_L$, the rank aggregation scheme should yield a unique, deterministic and total ranking of alternatives.

(S3) Transitivity. The aggregate ranking produced should be transitive, *ie.*
$\forall i, j, k \in \llbracket 1, n\rrbracket, (i \succ_F j) \text{ and } (j \succ_F k) \implies i \succ_F k$.

(S4) Pareto-efficiency. For every pair of alternatives i, j, if every voter prefers i over j then the aggregate ranking should prefer i over j, *ie.* $i \succ_F j$.

(IIA) Independence of irrelevant alternatives. The social preference between i and j should depend only on the individual preferences between i and j. *ie.* if one or more voters change their preferences, but no one changes their relative positions of i and j, then the relative positions of i and j in the aggregation should still remain unchanged.

Arrow's impossibility theorem. For $n > 2$, there doesn't exist any rank aggregation scheme satisfying the axioms (S1) - (S4), (IIA).

This theorem pertaining to the inconsistency of the previous axioms has been proved for ordinal rank aggregation schemes and is not valid for cardinal utility-based aggregation schemes. An axiomatic foundation of utility-based aggregation schemes entered the social choice theory literature with the work of Soufiani *et al.* (2014, [5]).

(S5) Anonymity. The rank aggregation scheme is insensitive to permutations over voters, *ie.* relabelling of the voters.

(S6) Monotonicity. If one of the voters moves an alternative i up in his ranking, the position of i in the aggregation can only improve.

(S7) Consistency. Consider two profiles of rankings $\mathcal{D}_L$ and $\mathcal{D}'_{L'}$ such that the aggregation scheme ranks the same alternative first in both, let's say i. Then, when fed with profile of rankings $\hat{\mathcal{D}} := \mathcal{D}_L \cup \mathcal{D}'_{L'}$ of size $L + L'$, the aggregation scheme should rank alternative i first.

(CC) Condorcet Criterion. If there exists an alternative preferred by the majority over all other alternatives then it must be ranked first, *ie.* $\forall i, j \in [\![1, n]\!], i \succ_M j \implies i \succ_F j$

1.5 Axiomatic analysis

Using the formalism of $F^* : \mathcal{P}_n \to \mathfrak{S}_n$, axioms (S1), (S3) and (S5) are automatically satisfied. The universality axiom (S2) simplifies as only unicity in the distributional framework. Pareto-efficiency (S4) writes as

$$\forall \pi \in \mathfrak{S}_n, p(\pi) > 0, \pi_i < \pi_j \implies i \succ_{F^*(p)} j.$$

Independence of irrelevant alternatives (IIA) can be formulated as

$$\forall i, j \in [\![1, n]\!], p, p' \in \mathcal{P}_n,\ p_{i \succ j} = p'_{i \succ j} \implies \operatorname{sign}\left(F^*_i(p) - F^*_j(p)\right) = \operatorname{sign}\left(F^*_i(p') - F^*_j(p')\right).$$

Monotonicity (S6) can be formalized by considering two voters $\pi, \pi' \in \mathfrak{S}_n$ such that $\pi'_i < \pi_i$ and $\forall j \neq i, \pi'_j \geq \pi_j$. For two densities $p, p' \in \mathcal{P}_n$ such that for a perturbation $\delta > 0$, we have

$$\begin{cases} p(\pi') = p'(\pi') + \delta \\ p(\pi) = p'(\pi) - \delta \\ \forall \pi'' \neq \pi, \pi', p(\pi'') = p'(\pi'') \end{cases}$$

then (S6) is satisfied if and only if $F^*(p)_i \leq F^*(p')_i$.

Consistency (S7) can be translated into the distributional framework by denoting p and p' the two densities corresponding to the two sets of preference information and $\hat{p} := \frac{p+p'}{2}$ the density corresponding to the union of the sets. Then (S7) is satisfied if and only if for all alternative i,

$$\left.\begin{array}{l} F^*(p)_i = 1 \\ F^*(p')_i = 1 \end{array}\right\} \implies F^*(\hat{p})_i = 1.$$

Finally the Condorcet Criterion in the distributional framework corresponds to the *majority rule*, *ie* if an alternative is ranked first by strictly more than half the voters then it must be ranked first in the aggregate. For $p \in \mathcal{P}_n$, (CC) writes as $\forall j \neq i,\ p_{i \succ j} > \frac{1}{2} \implies F^*(p)_i = 1$.

Prasad *et al.* used the characterization of the distributional rank aggregation schemes with loss functional to prove an impossibility theorem in the limiting case of the distributional framework. Every such scheme F^* can be written as

$$F^*(p) \in \arg\min_{\pi \in \mathfrak{S}_n} g(\pi, p) \quad \text{where} \quad g : \mathfrak{S}_n \times \mathcal{P}_n \to \mathbb{R}.$$

Prasad's impossibility theorem. For $n > 2$, there doesn't exist any distributional rank aggregation scheme satisfying simultaneously (S2) and (S4).

Arrow's theorem states that no aggregation procedure F can satisfy (S1) - (S4), (IIA). The corresponding theorem for F^* states that if $g(\pi, \cdot)$ is continuous with respect to the topology of $\mathcal{P}_n$, both universality and Pareto-efficiency cannot be satisfied simultaneously. Since Pareto-efficiency is a rather weak assumption to expect from an aggregation scheme, we must accept procedures returning multiple optimal aggregations.

1.6 Luce's Choice Axiom

Luce (2008, [6]) proposed to formalize the independence of irrelevant alternatives in the field of probabilities. For all $\mathcal{N} \subseteq [\![1, n]\!]$, we start by defining a probability density $P_{\mathcal{N}} : \mathcal{P}(\mathcal{N}) \to [0, 1]$ on the subsets of $\mathcal{N}$. We can then build a ranking model p by induction from P. For the subsets of alternatives $R \subseteq \mathcal{N} \subseteq [\![1, n]\!]$, we denote $P_{\mathcal{N}}(R)$ the probability that the chosen alternative in $\mathcal{N}$ is also in R.

Luce's Choice Axiom.

$$\forall R \subseteq \mathcal{N} \subseteq [\![1, n]\!],\ P_{[\![1,n]\!]}(R) = \begin{cases} P_{\mathcal{N}}(R) P_{[\![1,n]\!]}(\mathcal{N}) \text{ if } \forall i, j \in [\![1, n]\!], P_{\{i,j\}}(i) > 0 \\ P_{[\![1,n]\!] \setminus \{i\}}(\mathcal{N} \setminus \{i\}) \text{ otherwise.} \end{cases} \tag{LCA}$$

The first case is when no alternative can be excluded by P, otherwise, there exists $i, j \in [\![1, n]\!]$ such as $P_{\{i,j\}}(i) = 0$ and then the alternative i can be excluded by the probability function P. By taking $\mathcal{N} = \{i, j\}$ and equating the axiom for $R = \{i\}$ and $R = \{j\}$, we obtain the *odds ratio*

$$\frac{P_{\{i,j\}}(i)}{P_{\{i,j\}}(j)} = \frac{P_{[\![1,n]\!]}(i)}{P_{[\![1,n]\!]}(j)}.$$

Luce proved that the (LCA) is equivalent to the existence of a positive ratio scale $s \in \mathbb{R}^n$ acting as a probability on $[\![1, n]\!]$ when normalized. The marginal probability of selecting i first can be written as

$$\forall i \in [\![1, n]\!],\ p_{[\![1,n]\!]}(i) = \frac{s_i}{\sum_{j \in [\![1,n]\!]} s_j}.$$

This ratio, unique up to a multiplicative positive factor, is the motivation of utility-based aggregation schemes. Furthermore, the link with ranking models has been studied by Luce and Saaty, resulting in the following postulate which allows to compute the probability of a total ranking by induction $\pi = (i, \pi_{-i})$. We use the notation $p_{[\![1,n]\!]}(\pi)$ for the probability of the total ranking $\pi \in \mathfrak{S}_n$, we have

$$\begin{cases} \pi = (i, \pi_{-i}) \implies p_{[\![1,n]\!]}(\pi) = P_{[\![1,n]\!]}(i)\ p_{[\![1,n]\!] \setminus \{i\}}(\pi_{-i}) \\ p_{\{i,j\}}(i \succ j) = P_{\{i,j\}}(i). \end{cases}$$

The most famous model derived from the (LCA) is the Plackett-Luce model corresponding to score s. By construction, the aggregation schemes derived by maximizing such a model p satisfy the (IIA) axiom. Note that we usually omit the indice $[\![1, n]\!]$ since the alternative set is fixed.

2 Distance-based aggregation

Following the characterization of distributional rank aggregation schemes with loss functionals, we consider the class of procedures described by the bivariate loss functions $l : \mathfrak{S}_n \times \mathfrak{S}_n \to \mathbb{R}$ according to

$$F^*(p) = \underset{\pi \in \mathfrak{S}_n}{\arg\min}\, g(\pi, p) = \underset{\pi \in \mathfrak{S}_n}{\arg\min}\, \mathbb{E}_{\pi' \sim p}\big[l(\pi, \pi')\big]$$

- Machine Learning approach: corresponds to the $0-1$ loss function $l(h(x), y) = \mathbb{I}(h(x) \neq y)$. The approach tries to fit a unimodal probability distribution on the data and then returns the mode of the distribution as a point estimate.
- Distance-based approach: corresponds to a distance between rankings as loss. The choice of the loss function as a distance $\delta : \mathcal{R}^2 \to \mathbb{R}_+$ between rankings offers a wide class of schemes whose properties are determined by the choice of the distance. The corresponding optimization is

$$F^*_{\text{KT}}(p) = \underset{\pi \in \mathfrak{S}_n}{\arg\min}\, \mathbb{E}_{\pi' \sim p}\big[\delta(\pi, \pi')\big].$$

The function δ on rankings is a *metric* if it follows the following axioms.

(D1) Identity of indiscernibles. $\forall r, \tilde{r} \in \mathcal{R}, \delta(r, \tilde{r}) = 0 \iff r = \tilde{r}$.

(D2) Symmetry. $\forall r, \tilde{r} \in \mathcal{R}, \delta(r, \tilde{r}) = \delta(\tilde{r}, r)$.

(D3) Subadditivity. $\forall r, \tilde{r}, \hat{r} \in \mathcal{R}, \delta(r, \hat{r}) \leq \delta(r, \tilde{r}) + \delta(\tilde{r}, \hat{r})$.

From a metric δ between rankings, it is possible to build a distance between a ranking and a profile of partial rankings $\mathcal{D}_L$ by the simple rule $d(r, \mathcal{D}_L) := \sum_{l=1}^{L} \delta(r, \pi^{(l)})$. By searching for a condition when additivity holds, Kemeny and Snell proposed the following definition of *betweenness*. The ranking $\tilde{r}$ is *between* r and $\hat{r}$ if

$$\forall i, j \in [\![1, n]\!], \begin{cases} \tilde{r}_i < \tilde{r}_j \implies r_i < r_j \text{ or } \hat{r}_i < \hat{r}_j \\ \tilde{r}_i = \tilde{r}_j \implies (r_i - r_j)(\hat{r}_i - \hat{r}_j) \leq 0. \end{cases}$$

Kemeny and Snell proved that the *Kendall's* τ is the only metric between rankings satisfying all the following axioms. It can be defined on the set of rankings $\mathcal{R}$ as

$$\delta(r, \tilde{r}) = \sum_{i,j \in [\![1,n]\!]} \gamma_{i,j}(r, \tilde{r}) \text{ where } \gamma_{i,j}(r, \tilde{r}) := \begin{cases} 2 \text{ if } r_i < r_j \text{ and } \tilde{r}_i > \tilde{r}_j \\ 1 \text{ if } (r_i < r_j \text{ and } \tilde{r}_i = \tilde{r}_j) \text{ or } (r_i = r_j \text{ and } \tilde{r}_i > \tilde{r}_j) \\ 0 \text{ otherwise.} \end{cases}$$

(D4) Betweenness. If $\tilde{r}$ is between r and $\hat{r}$, then $\delta(r, \hat{r}) = \delta(r, \tilde{r}) + \delta(\tilde{r}, \hat{r})$.

(D5) Right-invariance. For all permutation $\pi \in \mathfrak{S}_n$, $\forall r, \tilde{r} \in \mathcal{R}, \delta(\pi(r), \pi(\tilde{r})) = \delta(r, \tilde{r})$.

(D6) Independence of extremal alternatives. If an alternative is added to $[\![1, n]\!]$ and ranked first or last by both r and $\tilde{r}$, then $\delta(r, \tilde{r})$ remains unchanged.

(D7) Normalization. The minimum distance is the unity, *ie.* $\min_{r \neq \tilde{r} \in \mathcal{R}} \delta(r, \tilde{r}) = 1$.

It corresponds to the usual formulation of the Kendall's τ between two permutations π and σ as the number of discordant pairs of alternatives

$$\tau_K(\pi, \tilde{\pi}) := \sum_{i,j \in [\![1,n]\!]} \mathbb{I}\Big((\pi_i > \pi_j) \wedge (\tilde{\pi}_i < \tilde{\pi}_j)\Big).$$

2.1 Kemeny orders

Assuming the existence of an underlying *ground truth*, *ie.* a true total ranking π_{GT} depicting an absolute truth, Condorcet addressed the question of the most likely aggregated ranking given a profile of rankings $\mathcal{D}_L$. He modeled the pairwise choice of voters with a probability $q \in]\frac{1}{2}, 1]$ to make the right choice. Young (1988) completed this statistical approach with a link between the Kemeny distance and the maximum likelihood.

Let's denote by $\succ_{GT}$ the total order associated with the ground truth π_{GT}. Knowing the ground truth, the probability of observing a profile $\mathcal{D}_L = \{\pi^{(1)}, \cdots, \pi^{(L)}\}$ is

$$p(\mathcal{D}_L \mid \pi_{GT}) = q^{\sum_{i \succ_{GT} j} v_{i,j}(\mathcal{D}_L)} (1-q)^{\sum_{i \succ_{GT} j} v_{j,i}(\mathcal{D}_L)}.$$

The maximum likelihood problem corresponds to estimating the parameter π^* which maximizes the likelihood $\mathcal{L}(\pi \mid \mathcal{D}_L) = p(\mathcal{D}_L \mid \pi)$. It corresponds to maximizing the number of pairwise support for a given π in the profile $\mathcal{D}_L$. We denote π_L^* one of the optima of this problem,

$$\pi_L^* \in \arg\max_{\pi \in \mathfrak{S}_n} \sum_{\substack{i,j \in [\![1,n]\!] \\ \pi_i < \pi_j}} v_{i,j}(\mathcal{D}_L).$$

Condorcet claimed that as the number of pairwise preferences tends to the infinity, the recovered ranking corresponds to the ground truth π_{GT}. In our listwise framework, the claim corresponds to the convergence in the sense of Kendall

$$\tau_K(\pi_L^*, \pi_{GT}) \xrightarrow[L \to \infty]{} 0.$$

The *Kemeny distance* is defined as the sum of each Kendall's τ between the given permutation and a ranking of the profile $d_K(\pi, \mathcal{D}_L) := \sum_{l=1}^{L} \tau_K(\pi, \pi^{(l)})$. By permuting the elements in the sums, the maximization problem can be cast into a Kemeny distance minimization

$$F_{\mathrm{KT}}(\mathcal{D}_L) \in \arg\min_{\pi \in \mathfrak{S}_n} d_K(\pi, \mathcal{D}_L).$$

One such $F_{\mathrm{KT}}(\mathcal{D}_L) \in \mathcal{R}$ is called a *Kemeny order*. The problem has been shown to be NP-complete even for four votes (1989, [7]) which motivates the use of specific stochastic optimization algorithms optimizing with respect to this Kemeny criterion, such as the Cross Entropy Monte Carlo (CEMC) algorithm.

Furthermore, if a Condorcet winner exists then it is ranked first. (XCC) is also satisfied by any Kemeny order but unfortunately, checking the (XCC) is intractable since it requires enumerating the B_n partitions in $\mathcal{P}_0([\![1, n]\!])$ where B_n is the number of Bell, a growth rate worst than exponential since $B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} B_k$. Truchon proposed an algorithm exhibiting several bins of the partition, particularly useful for testing purposes: a proposed order cannot be Kemeny if the singletons don't correspond or if the cycle bins of the *ex æquo* relation $=_M$ are not the same up to a permutation.

2.2 Cross Entropy Monte Carlo method

The Cross Entropy method has initially been designed to simulate rare events since it iteratively improves a probability distribution to estimate the probability of a given *rare-event*. A Monte Carlo procedure is used to sample solutions that are more likely to correspond to the rare-event as the number of iteration grows and the probability distribution improves. This Monte Carlo procedure can be formalized for combinatorial optimization by considering the rare-event *being close to an optima*. The method was initially used to estimate rare event probabilities with an objective of the form $l(\gamma) = \mathbb{P}(S(X) \leq \gamma)$ where $S(X) \leq \gamma$ is the rare event corresponding to a *small enough* value of S.

The distance-based aggregation with the Kendall's τ is $F_{\mathrm{KT}}^*(p) = \arg\min_{\pi \in \mathfrak{S}_n} \mathbb{E}_{\pi' \sim p}\big[\tau_K(\pi, \pi')\big]$.

In the case of finite available information $\mathcal{D}_L$, the population problem can be approximated with the usual rank aggregation scheme of Kemeny

$$F_{\text{KT}}(\mathcal{D}_L) = \operatorname*{arg\,min}_{\pi \in \mathfrak{S}_n} \sum_{l=1}^{L} \tau_K(\pi, \pi^{(L)}) = \operatorname*{arg\,min}_{\pi \in \mathfrak{S}_n} d_K(\pi, \mathcal{D}_L).$$

The Spearman footrule is defined with the first order metric $\delta_F(\pi, \pi') := \sum_{i \in [\![1,n]\!]} |\pi_i - \pi'_i|$ by summing over the profile, $d_F(\pi, \mathcal{D}_L) := \sum_{l=1}^{L} \delta_F(\pi, \pi^{(l)})$. It is often used to approximate the Kemeny distance since we have the 2-approximation for all $\pi \in \mathcal{R}, d_K(\pi, \mathcal{D}_L) \leq d_F(\pi, \mathcal{D}_L) \leq 2d_K(\pi, \mathcal{D}_L)$. The distributional Spearman footrule

$$F^*_{\text{SF}}(p) = \operatorname*{arg\,min}_{\pi \in \mathfrak{S}_n} \mathbb{E}_{\pi' \sim p} \sum_{i \in [\![1,n]\!]} |\pi_i - \pi'_i|$$

becomes the Spearman footrule, which makes the optimization tractable

$$F_{\text{SF}}(\mathcal{D}_L) = \operatorname*{arg\,min}_{\pi \in \mathfrak{S}_n} d_F(\pi, \mathcal{D}_L).$$

2.3 Stochastic optimization

We consider the Kemeny optimization problem for a finite set of preferences $\mathcal{D}_L$. Therefore, the goal is to minimize $d_K(\pi, \mathcal{D}_L)$ over the set of permutations $\pi \in \mathfrak{S}_n$ where $d_K(\cdot, \mathcal{D}_L)$ is the Kemeny distance against the profile of rankings $\mathcal{D}_L$. A global minima is such that

$$\gamma^* := d_K(\pi^*, \mathcal{D}_L) = \min_{\pi \in \mathfrak{S}_n} d_K(\pi, \mathcal{D}_L).$$

We randomize the deterministic problem into the so-called *associated stochastic problem* (ASP). Since the set of permutations is finite, we can write

$$l(\gamma) = \mathbb{P}(d_K(\pi, \mathcal{D}_L) \leq \gamma) = \mathbb{E}[\mathbf{1}_{d_K(\pi, \mathcal{D}_L) \leq \gamma}] = \sum_{\pi} \mathbf{1}_{d_K(\pi, \mathcal{D}_L) \leq \gamma} f(\pi) \tag{ASP}$$

where f is the real distribution. The rare-event estimation corresponds to the estimation of $l(\gamma)$ for a given threshold γ. Another associated problem would be to estimate the threshold γ for a given probability $l(\gamma)$, *ie.* estimate the root of the (ASP). Optimization can be achieved by considering the first problem of estimating the probability $l(\gamma)$ for a certain γ close to the optimal value γ^*.

The main idea of the Cross Entropy method is to iteratively approximate f by a parameterized distribution f_p by improving the parameter p. We initially chose f_u as a prior distribution over the permutations.

By using *importance sampling*, we define a new probability distribution g to write

$$\sum_{\pi} \mathbf{1}_{d_K(\pi, \mathcal{D}_L) \leq \gamma} \frac{f_u(\pi)}{g(\pi)} g(\pi) = \mathbb{E}_g \left[\mathbf{1}_{d_K(\pi, \mathcal{D}_L) \leq \gamma} \frac{f_u(\pi)}{g(\pi)} \right].$$

The probability distribution g minimizing the Cross Entropy is $g^*(\pi) := \frac{1}{l(\gamma)} f_u(\pi) \mathbf{1}_{d_K(\pi, \mathcal{D}_L) \leq \gamma}$ but requires the knowledge of $l(\gamma)$, the quantity of interest. In the Cross Entropy method, we look for g in the same family as the prior f_u by choosing $g := f_{p^*}$ with the parameter p^* which minimizes the Cross Entropy between the two distributions g^* and f_p. We denote D_{KL} the Kullback-Leibler divergence, *ie.* the Cross Entropy between two finite distributions. We have

$$\begin{aligned}
p^* &= \arg\min_p D_{KL}(g^*, f_p) \\
&= \arg\min_p \sum_\pi g^*(\pi)\big(\log g^*(\pi) - \log f_p(\pi)\big) \\
&= \arg\max_p \sum_\pi \log f_p(\pi) g^*(\pi) \\
&= \arg\max_p \sum_\pi \mathbf{1}_{d_K(\pi,\mathcal{D}_L)\le\gamma} \log f_p(\pi) f_u(\pi)
\end{aligned}$$

If $\pi_1, \cdots, \pi_N \sim^{iid} f_u$ are samples, the maximization problem can be approximated using the Monte Carlo estimator, which is unbiased. The corresponding problem, called the *stochastic counterpart*, writes as

$$\hat{p^*} = \arg\max_p \frac{1}{N} \sum_{k=1}^{N} \mathbf{1}_{d_K(\pi,\mathcal{D}_L)\le\gamma} \log f_p(\pi_k). \tag{SC}$$

We can build the CE procedure by iteratively solving this stochastic counterpart problem to improve the parameterized distribution f_p. The main limitation lies in the choice of the probability distribution family which determines the number of degrees of freedom. In [8], Rubinstein *et al.* stated that in the case of a unique optimizer π^* corresponding to the optimal value γ^*, if the class of distribution $\{f_p \mid p\}$ contains the *atomic* density with mass at π^*

$$\delta_{\pi^*}(\pi) = \begin{cases} 1 & \text{if } \pi = \pi^* \\ 0 & \text{otherwise} \end{cases}$$

then the solution of the optimization is the density δ_{π^*}. Note that the unicity is theoretically ensured by enforcing an artificial order on the set $\mathfrak{S}_n$. We can define a modified objective $S(\pi) = d_K(\pi, \mathcal{D}_L) - \epsilon(\pi)$ where $\epsilon(\pi)$ is a small perturbation breaking the ties. We can choose $\epsilon(\pi)$ proportional to the rank of π in the defined order and small enough to ensure that $S(\pi) > S(\pi')$ if $d_K(\pi, \mathcal{D}_L) > d_K(\pi', \mathcal{D}_L)$. Therefore the solution of the optimization with the perturbed objective S is the atomic density with all the mass on the optimizer with the highest ranking.

Furthermore, if f is chosen in the exponential family, the estimator $\hat{p^*}$ has a closed form which corresponds to the usual Maximum Likelihood Estimator with the addition of the indicator functions (2013, [9]).

2.4 Multi-level formulation

Unlike the rare event simulation where the threshold γ is fixed by the definition of the event itself, in the adaptive uploading algorithm we build a sequence of $\hat{\gamma_t}$ along with the sequence of parameters $\hat{p_t}$. The algorithm is called *multi-level* since the objective event $d_K(\pi, \mathcal{D}_L) \le \hat{\gamma_t}$ improves along with the parameter.

For a fixed p_{t-1}, we denote γ_t the $(1-\rho)$-quantile of the random variable $S(\pi)$ under p_{t-1}. This quantile can be estimated using the usual sample estimator $\hat{\gamma_t} = S_{(\lceil(1-\rho)N\rceil)}$ where S is a vector of the Kemeny distances $S_1 = d_K(\pi_1, \mathcal{D}_L), \cdots, S_N = d_K(\pi_N, \mathcal{D}_L)$ with the N samples $\pi_1, \cdots, \pi_N \sim^{iid} f_{p_{t-1}}$.

Algorithm 1 Multi-level Cross Entropy Monte Carlo

```
procedure MultiLevelCemc($u, \rho, \delta, N, N'$)
    $p_0 \leftarrow u$                                          ▷ initialize with prior parameter
    $t \leftarrow 1$
    loop
        generate $N$ samples $\pi_1, \cdots, \pi_N \sim^{iid} f_{p_{t-1}}$
        compute the Kemeny distances $S_1 = d_K(\pi_1, \mathcal{D}_L), \cdots, S_N = d_K(\pi_N, \mathcal{D}_L)$
        $\hat{\gamma_t} \leftarrow S_{(\lceil (1-\rho)N \rceil)}$
        denote $p_t$ the solution of (SC) with *same* samples $\pi_1, \cdots, \pi_N$
        $t \leftarrow t + 1$
    until ($\widehat{\gamma_{t-\delta}} = \cdots = \widehat{\gamma_t}$)
    $T \leftarrow t$
    $\hat{l} := \frac{1}{N'} \sum_{i=1}^{N'} \mathbf{1}_{d_K(\pi, \mathcal{D}_L) \leq \gamma} \log f_{p_T}(\pi_k)$ for $\pi_1, \cdots, \pi_{N'} \sim^{iid} f_{p_T}$
    return $\hat{l}$                    ▷ return a Monte Carlo estimation with more samples $N' > N$
```

The algorithm uses the information of the prior u and depends on the convergence parameters N, ρ and δ. The parameter $\delta \in \mathbb{N}$ controls the number of steps without improvement before the stopping criterion is reached. Kroese notes that the algorithm is *self-tuning* for the other parameters.

In the population setting, the multi-level Cross Entropy Monte Carlo algorithm corresponds to the following version.

Algorithm 2 Multi-level Cross Entropy with population setting

```
procedure MultiLevelCeDeterministic($u, \rho, \delta, N, N'$)
    $p_0 \leftarrow u$                                          ▷ initialize with prior parameter
    $t \leftarrow 1$
    loop
        $\gamma_t \leftarrow \max\{\gamma \mid \mathbb{P}_{p_{t-1}}(d_K(\Pi, \mathcal{D}_L) \leq \gamma) \geq \rho\}$
        $p_t \leftarrow \arg\max_p \mathbb{E}_{p_{t-1}} \mathbf{1}_{d_K(\Pi, \mathcal{D}_L) \leq \gamma_t} \log f_p(\Pi)$
        $t \leftarrow t + 1$
    until ($\gamma_{t-\delta} = \cdots = \gamma_t$)
    $T \leftarrow t$
    $l := \mathbb{P}_{p_T}(d_K(\Pi, \mathcal{D}_L) \leq \gamma_T)$
    return $l$                                         ▷ return the estimated probability
```

In the Order Explicit Algorithm (OEA) case (2010, [10]), the parameter p is taken as a matrix of size $n \times n$ where the i$^{\text{th}}$ column specifies the probability distribution over the n alternatives for the i$^{\text{th}}$ position in the aggregate. The OEA stores the probability distribution of the alternative at each position of the permutation in a matrix $p = ((p_{i,j}))$ which acts as a parameter of the distribution f_p. Sampling from these specifications requires a specific procedure used in the Monte Carlo version.

2.5 Local convergence

Even if the OEA specifications have a sufficient expressiveness to contain the atomic density corresponding to one optimizer, the Cross Entropy Monte Carlo method itself is sensitive to the problem of local convergence and might converge to a local optimum. Furthermore, the notion of locality on the set of permutations $\mathfrak{S}_n$ depends on the metric considered. The devised algorithm optimizes in the metric space $\left(\mathfrak{S}_n, \tau_K\right)$ where τ_K is the Kendall's τ. The algorithm with the OEA encoding proceeds to *blur* the parameter matrix to explore the surroundings in the sense of swapping adjacent alternatives. From the parameter p we define the blurred matrix p' with blurring parameter $b \in [0, \frac{1}{2}]$ as

$$\forall i \in \{1, \cdots, n\}, p'_{i,j} = \begin{cases} (1-b)p_{i,j} + bp_{i,j+1} & \text{if } j = 1 \\ bp_{i,j-1} + (1-b)p_{i,j} & \text{if } j = n \\ bp_{i,j-1} + (1-2b)p_{i,j} + bp_{i,j+1} & \text{otherwise} \end{cases}$$

This transformation preserves the columns summing to 1, hence p' also characterizes a $k \times n$ degrees of freedom distribution over the permutations.

$$p = \begin{bmatrix} p_{1,1} & \cdots & p_{1,k} \\ \vdots & & \vdots \\ p_{n,1} & \cdots & p_{n,k} \end{bmatrix} \xrightarrow{\text{blurring}} p' = \begin{bmatrix} \cdots & bp_{1,j-1} + (1-2b)p_{1,j} + bp_{1,j+1} & \cdots \\ & \vdots & \\ \cdots & bp_{n,j-1} + (1-2b)p_{n,j} + bp_{n,j+1} & \cdots \end{bmatrix}$$

Formally, we consider that two permutations are close if it only requires a few swaps to go from one to the other. This is formalized by balls in the metric space $(\mathfrak{S}_n, \tau_K)$. In this space, we can define balls centered on a permutation σ as $B_r(\sigma) = \{\pi \in \mathfrak{S}_n \mid \tau_K(\pi, \sigma) \leq r\}$ useful for testing the convergence.

Let's denote $C_\sigma(k, r)$ the set of permutations $\pi \in \mathfrak{S}_n$ such that $\tau_K(\pi, \sigma) \leq r$ and only the top-k elements of σ are not determined *ie.* the positions of the other elements are fixed. The following decomposition is based on the fact that inserting the k^{th} alternative of σ at position i creates $k - i$ new inversions, so

$$\begin{aligned} C_\sigma(k, r) &= \bigsqcup_{i \mid k-i \leq r} C_{p_{k,i}(\sigma)}\big(k-1, r-(k-i)\big) \\ &= \bigsqcup_{i=\max(1,k-r)}^{k} C_{p_{k,i}(\sigma)}\big(k-1, r-k+i\big). \end{aligned}$$

where $p_{k,i}(\sigma)$ denotes the permutation obtained by inserting the k^{th} element of σ at position i. The base cases of the recursive definition are $C_\sigma(0, \cdot) = \emptyset$ and $C_\sigma(\cdot, 0) = \{\sigma\}$. When applied recursively, this procedure explores the closest permutations in a tree-like manner by pruning the useless calls, *ie.* those exhibiting permutations that are not in the desired ball.

This procedure recovers the desired ball since $B_r(\sigma) = C_\sigma(n, r)$. The case $r = \infty$ illustrates a constructive and recursive definition of $\mathfrak{S}_n$ since inserting $\{n\}$ at any position in $(1, \cdots, n-1)$ and then recursively doing the same with $n-1$ among the remaining elements yields the entire set of permutations, *ie.* $B_\infty(\sigma)$.

Note that dynamic programming reduces the runtime complexity of the implementation at the cost of additional spatial complexity. The size of the ball $B_r(\sigma)$ is also exponential in r. The CEMC algorithm converges to a local minimum for the Kemeny distance in the sense of the Kendall's Tau ball, *ie.* if π^∞ is the limit ranking produced, it also has the smallest Kemeny distance among the rankings of $B_r(\pi^\infty)$ for r small enough.

2.6 Sampling permutations

The Monte Carlo method heavily relies on two elemental operations: sampling a permutation from the parameterized distribution and computing the Kemeny score, *ie.* the *Kemenization* of each permutation. therefore, the sampling procedure constitutes a bottleneck in the algorithm which must be efficiently implemented. We propose to investigate how to efficiently sample permutations from the OEA specification of the CEMC algorithm. Note that sampling and Kemenization procedures are easily parallelizable and that the only sequential operation of a CEMC iteration is computing the quantile to retain the elite samples.

To sample from the OEA matrix $p = ((p_{i,j}))$ with $i \in [\![1, n]\!],\ j \in [\![1, k]\!]$, the naive method is to apply an Inverse Transform Sampling (ITS) for each alternative from top to bottom. We sample the top-k

alternatives starting from the 1st column by applying successive ITS. For each $j \in \{1, \cdots, k\}$, we use the j^{th} column as the discrete probability distribution from which the j^{th} alternative is sampled. For $j > 1$, we need to cancel out the probabilities corresponding to already chosen alternatives and renormalize the whole vector which is expensive. The total cost of sampling a permutation from p is $\mathcal{O}(n^2)$.

Walker (1974, [11]) proposed a method to improve the performance of sampling from a discrete distribution q to a constant time $\mathcal{O}(1)$ after a $\mathcal{O}(n)$ preprocessing. Graphically, the main idea can be seen on the histogram of q where each alternative i is represented by a bin whose area is q_i. We uniformly sample a point (x, y) of the plane to find the sampled alternative i directly on the histogram. To avoid the high rate of rejection associated with the *blank space*, when (x, y) falls off a bin on the histogram, the mass exceeding $\frac{1}{n}$ of the *overfull bins* is redistributed to the *underfull bins* in such a way that the bin corresponding to (x, y) can be determined in constant time. This representation is called a *binary mixture* since each abscisse index contains at most 2 bins.

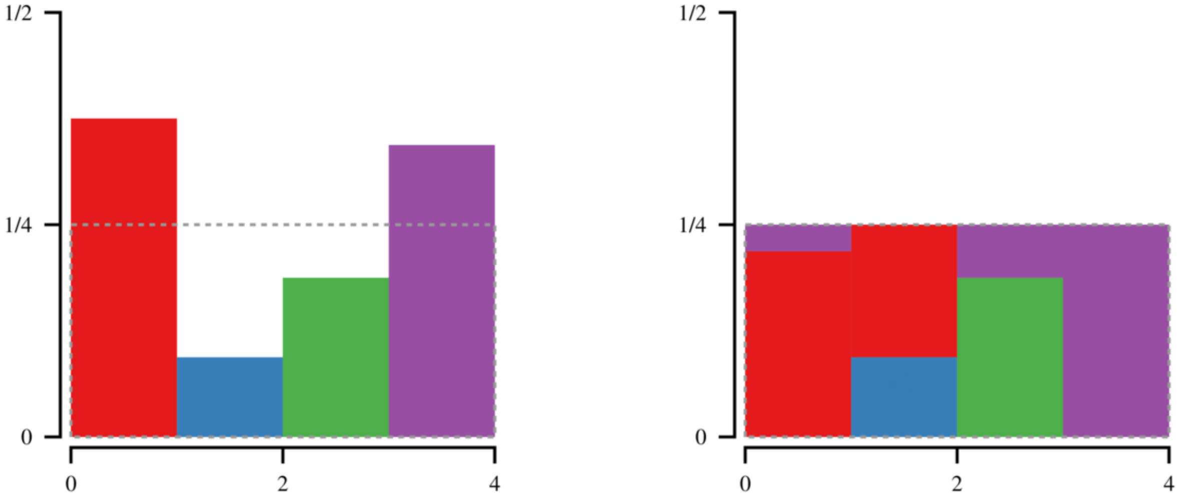


Figure 1: Alias method applied on a toy example of discrete distribution $q = (q_1, q_2, q_3, q_4)$. On the left the histogram and on the right an associated binary mixture of surface 1 representing the same distribution on the plane.

The existence of a binary mixture representation is insured by the following theorem due to Devroye (1986, [12]). Every probability vector q can be expressed as an equiprobable mixture of two-points distributions, *ie.*

$$\exists (i_0, j_0), \cdots, (i_n, j_n) \in \mathbb{N}^{2n}, (\tilde{q}_0, \cdots, \tilde{q}_n) \in [0, 1]^n, \sum_{i=1}^{n} \tilde{q}_i = 1,$$

$$\forall i \in \{1, \cdots, n\}, q_i = \frac{1}{n} \sum_{l=1}^{n} \left(\tilde{q}_l \mathbb{I}(i_l = i) + (1 - \tilde{q}_l) \mathbb{I}(j_l = i) \right)$$

Note that the representation is not unique. The Alias method can be extended to sample permutations by first preprocessing a binary mixture of each column p_j. Then an alternative from any p_j can be sampled in constant time which allows the use of a rejection procedure to sample a permutation. Sampling an alternative at position j close to bottom (*ie.* j close to k) will cost more rejections than j close to the top.

The Monte Carlo algorithm requires sampling N permutations from the distribution parameterized by p which costs $\mathcal{O}(N \times n^2)$ with the naive method against a cost between $\mathcal{O}(N \times n)$ and $\mathcal{O}(N \times n^2)$ with the Alias method. The worst case corresponds to a high rate of rejection, but in practice the Alias method is more efficient than the naive method since it breaks the square complexity.

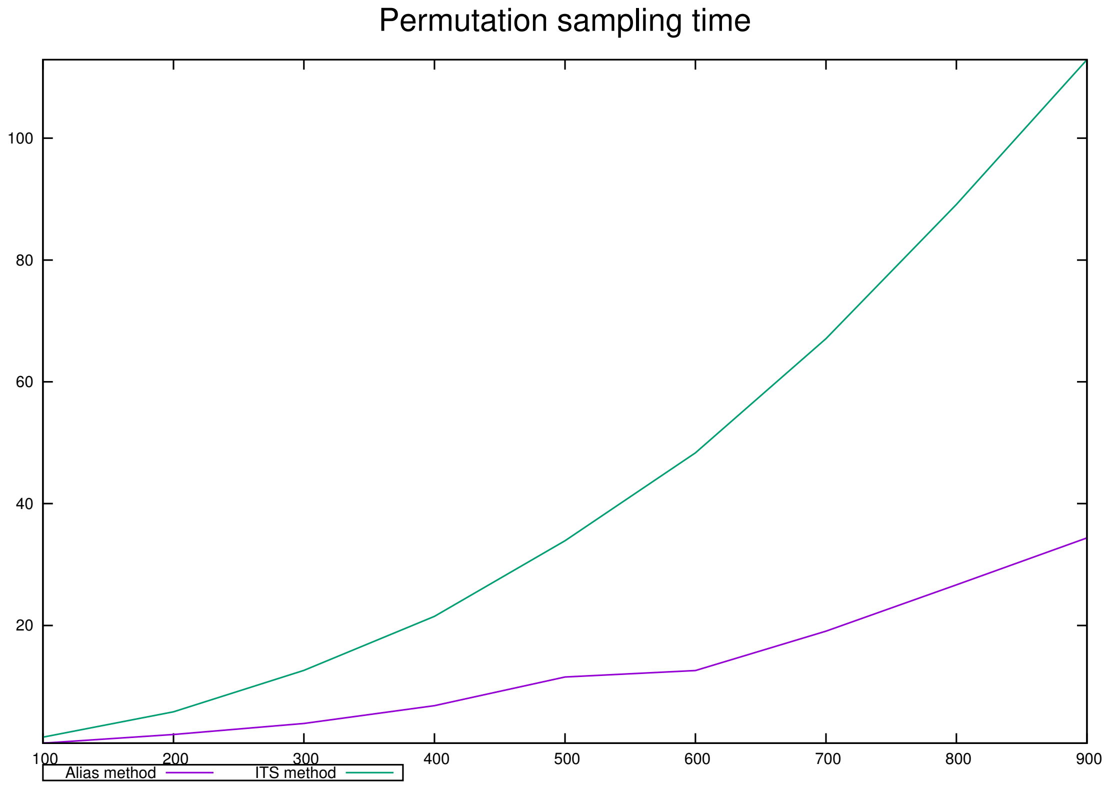


Figure 2: Time required (in seconds) to sample $N = 10,000$ permutations from a given non-trivial OEA matrix p with the ITS method and with the Alias method as a function of the number of alternatives n.

3 Partial rankings and marginals

3.1 Ranking model

Some performance measures being computationally challenging to evaluate for non-linear investment strategies, *eg.* requiring the use of Monte Carlo simulations, we would like to leverage on *partial ranking information* where only a subset of the alternatives are ranked. This kind of incomplete data naturally arises in the Information Retrieval field where not all comparisons are provided to the aggregation algorithms. By denoting x an alternative, the most studied partial information rankings, the *bucket orders* are of the form

$$x_{1,1}, \cdots, x_{1,n_1} \succ \cdots \succ x_{r,1}, \cdots, x_{r,n_r} \text{ where } r \leq n \text{ and } \sum_{i=1}^{r} n_i = n.$$

This corresponds to the case studied by Truchon with the Extended Condorcet Criterion with partition $X = \{X_1, \cdots, X_r\} \in \mathcal{P}_0(\llbracket 1, n \rrbracket)$ with $\forall i \leq r, X_i = \{x_{i,1}, \cdots, x_{i,n_i}\}$. Bucket orders also generalize the usual top-k rankings where only the best $k \leq n$ alternatives are ranked by considering $x_1 \succ x_2 \succ \cdots \succ x_k \succ x_{k+1}, \cdots, x_n$. Clémençon *et al.* (2014,[13]) introduced a Multiresolution Analysis (MRA) on the space of functionals of $\mathfrak{S}_n$ to formalize the idea of learning a ranking model by estimating the marginals.

To formally define a distribution over partial rankings, we use the concept of *injective words* made of different alternatives $\pi = \pi_1 \cdots \pi_k$ and of size $|\pi| = k$. We write $\pi' \subseteq \pi$ if there exists indices $i_1 < \cdots < i_{|\pi'|}$ such as $\pi' = \pi_{i_1} \cdots \pi_{i_{|\pi'|}}$. We denote by $\Gamma(\mathcal{N})$ the set of injective words on the subset $\mathcal{N} \subseteq \llbracket 1, n \rrbracket$. All the ranking information belongs to the union set

$$\Gamma_n := \bigcup_{\mathcal{N} \subseteq \llbracket 1, n \rrbracket} \Gamma(\mathcal{N}).$$

For a subset $\mathcal{N} \subseteq \llbracket 1, n \rrbracket$, the marginal probability distribution of $\mathcal{N}$ over $\Gamma(\mathcal{N})$ is defined by

$$\forall \pi \in \Gamma(\mathcal{N}),\ p_{\mathcal{N}}(\pi) := \sum_{\substack{\sigma \in \mathfrak{S}_n \\ \pi \subseteq \sigma}} p(\sigma).$$

We define a probability density ν on the power set $2^{\llbracket 1, n \rrbracket}$ to draw the subset spaces $\mathcal{N}$. We restrict the study to the support of ν *ie.* the non-null subsets $\text{supp}(\nu) = \{\mathcal{N} \subseteq \llbracket 1, n \rrbracket \mid \nu(\mathcal{N}) > 0\}$ of size $Q := |\text{supp}(\nu)|$. Each subset $\mathcal{N}$ defines a batch of partial ranking data $\{\pi^{(l,\mathcal{N})}\}_{l=1}^{L^{(\mathcal{N})}}$. If the distributional assumption holds, then the following marginal distributional assumption also holds by taking the marginals of the global ranking model p.

Marginal distributional assumption. There exists distributions $p_{\mathcal{N}}$ such that

$$\forall \mathcal{N} \in \text{supp}(\nu), \begin{cases} \mathcal{N} \sim \ \nu \\ \forall l \in \{1, \cdots, L^{(\mathcal{N})}\},\ \pi^{(l,\mathcal{N})} \sim^{iid} p_{\mathcal{N}}. \end{cases}$$

3.2 Identifiability

Without structural assumptions on the ranking model p, it has $n! - 1$ degrees of freedom which makes its identifiability intractable. For example, the pairwise setup which stores preference judgments into a matrix J only provides access to the pairwise marginals $p_{i \succ j}$ and therefore only $n(n-1)/2$ parameters of p are identifiable. The total number of accessible parameters in the partial ranking framework, *ie.* the number of parameters that can be learned is

$$\sum_{\mathcal{N} \in \text{supp}(\nu)} (|\mathcal{N}|! - 1).$$

Using the multiresolution analysis on partial rankings, Clémençon *et al.* ([13]) derived an identifiability result in the case of the partial ranking dataset. We denote by $!k$ the subfactorial function applied in k, *ie.* the number of derangements on a set of size k. The derangements of a set are the fixed-point free permutations, *ie.* the permutations such that no element remains in its original position.

Clémençon's identifiability theorem. In the absence of restrictive assumption on the ranking model p, only the marginals $p_{\mathcal{N}}$ for $\mathcal{N} \in \mathcal{P}(\text{supp}(\nu))$ are identifiable and characterized by $\sum_{\mathcal{N}\in\mathcal{P}(\text{supp}(\nu))} !|\mathcal{N}|$ degrees of freedom.

The number of derangements $!k$ is growing *almost as fast* as the factorial $k!$ since at the limit $!k =_{k\to\infty} \frac{k!}{e}$. This motivates the use of a structural assumption, *ie.* a model on $\mathfrak{S}_n$ or $\mathcal{R}$. The φ-model proposed by Mallows constitutes a broad class of models depending on a distance between permutations.

In the case of a global structure $\mathcal{M}(\theta)$ with parameter $\theta \in \Theta$, we use the following definition of identifiability. The statistical model is identifiable if

$$\forall\theta,\theta' \in \Theta, \forall\pi \in \mathfrak{S}_n,\ p_{\mathcal{M}}(\pi \mid \theta) = p_{\mathcal{M}}(\pi \mid \theta') \implies \theta = \theta'.$$

3.3 Condorcet model

In the case where $\succeq_M$ reduces to $\succ_M$ and contains no cycles, the aggregation problem with the majority relation consists into uncovering the finest partition and a final total ranking $F(\mathcal{D}_L)$ satisfying (XCC). Truchon (1998, [14]) proposed an algorithm to exhibit such a partition in a polynomial time $\mathcal{O}(n^3)$. When the relation $\succeq_M$ contains cycles, (XCC) doesn't address the issue of ranking inside the bins. In this case, the proposed algorithm produces a partial ranking in $\mathcal{R}$ with a *residual* bin whose alternatives remain to be ranked with another method.

Following the assumption of Condorcet, *ie.* independence of pairwise preferences, we consider that each voter provides a partial ranking $\pi^{(l)} \in \mathcal{R}$. The model depends on a ground truth $\pi_{GT} \in \mathcal{R}$ and is characterized by a non-decreasing function $q : [\![1, n-1]\!] \to]\frac{1}{2}, 1[$ which generalizes the constant q used to introduce the Kemeny orders. The pairwise marginals write as

$$\begin{array}{rcl} p_{\text{C}}(i \succ j \mid \pi_{GT}) & = & q(\pi_j - \pi_i) \\ p_{\text{C}}(j \succ i \mid \pi_{GT}) & = & 1 - q(\pi_j - \pi_i). \end{array}$$

Following the work of Drissi *et al.* (2002, [15]), we aggregate the pairwise preferences under the assumption of independence with the random variable associated to the marginal weight $V_{i,j}$. We have

$$p_{\text{C}}(V_{i,j} = u \mid \pi_{GT}) = \binom{L}{u}\ p_{\text{C}}(i \succ j \mid \pi_{GT})^u\ p_{\text{C}}(j \succ i \mid \pi_{GT})^{L-u}.$$

Therefore, the likelihood of the data $\mathcal{D}_L$ in the Condorcet model with function q is

$$\mathcal{L}(\pi \mid \mathcal{D}_L) = p_{\text{C}}(\mathcal{D}_L \mid \pi_{GT}) = \prod_{\substack{i,j\in[\![1,n]\!] \\ \pi_i<\pi_j}} p_{\text{C}}(V_{i,j} = v_{i,j}(\mathcal{D}_L) \mid \pi_{GT}).$$

By taking a logarithmic transformation and then rearranging the pairwise preferences by layers of distance

$\pi_j - \pi_i = k$ with $k \in [\![1, n-1]\!]$, we have

$$\begin{aligned}\operatorname*{arg\,max}_{\pi \in \mathcal{R}} \log \mathcal{L}(\pi \mid \mathcal{D}_L) &= \operatorname*{arg\,max}_{\pi \in \mathcal{R}} \sum_{\substack{i,j \in [\![1,n]\!] \\ \pi_i < \pi_j}} \log \Big(q(\pi_j - \pi_i)^{v_{i,j}(\mathcal{D}_L)} \big(1 - q(\pi_j - \pi_i)\big)^{L - v_{i,j}(\mathcal{D}_L)} \Big) \\ &= \operatorname*{arg\,max}_{\pi \in \mathcal{R}} \sum_{k=1}^{n-1} \sum_{\substack{i,j \in [\![1,n]\!] \\ \pi_j - \pi_i = k}} \Big(v_{i,j}(\mathcal{D}_L) \log q(k) + (L - v_{i,j}(\mathcal{D}_L)) \log(1 - q(k)) \Big) \\ &= \operatorname*{arg\,max}_{\pi \in \mathcal{R}} \sum_{k=1}^{n-1} \log \frac{q(k)}{1 - q(k)} \sum_{\substack{i,j \in [\![1,n]\!] \\ \pi_j - \pi_i = k}} v_{i,j}(\mathcal{D}_L)\end{aligned}$$

The maximum likelihood approach gives a class of aggregation schemes by varying the non-decreasing function q, which only appears as a weight vector $(\log \frac{q(k)}{1-q(k)})_{k=1}^{n-1}$ depending on k. The corresponding scheme is

$$F_q(\mathcal{D}_L) \in \operatorname*{arg\,max}_{\pi \in \mathcal{R}} \ \log \mathcal{L}(\pi \mid \mathcal{D}_L).$$

By considering the specific case where $q(k) = q \in]\frac{1}{2}, 1[$ is constant, we recover the result of Young, *ie.* the aggregation scheme corresponds to the Kemeny rule F_{KT}. Just like F_{KT}, the schemes F_q satisfy the axioms of non-dictatorship (S1) and anonymity (S5) but they are not Pareto-efficient (S4) nor satisfy (IIA). Drissi *et al.* showed that for all non-decreasing function q, the scheme F_q satisfies weaker versions of those two axioms, the *Weak Pareto Principle* and the *local independence of irrelevant alternatives.*

(WPP) Weak Pareto Principle. For an aggregation rule $F(\mathcal{D}_L)$, we call i the alternative ranked first, *ie.* such as $F(\mathcal{D}_L)_i = 1$ and $\nexists j \in [\![1,n]\!], \forall l \ in [\![1, L]\!], j \succ_{\pi^{(l)}} i$.

This weaker version of the Pareto-effiency only concerns the alternative ranked first. Axiomatic results for the top-k alternatives of an aggregation scheme is more challenging. The utility-based aggregation theory circumvents the impossibility theorems by building schemes of the choice axiom of Luce (LCA). The impossibility theorem of Arrow excludes the possibility of finding non-trivial paretian aggregation schemes satisfying the independence of irrelevant alternatives, which led Yound and Levenglick (1977, [16]) to propose a weakened version of the (IIA) axiom.

(LIIA) Local Independence of Irrelevant Alternatives. If a subset of the alternatives are in consecutive positions in the aggregation, then their relative order must remain unchanged if all other alternatives are deleted from the profile.

3.4 Mallows model

The Condorcet model is built *bottom up*, from the pairwise marginals to rankings and not directly on the rankings themselves. On the other hand, Mallows (1957, [17]) proposed a φ-model $\mathcal{M}(\pi, \varphi)$ as a permutation distribution allowing a wide range of metrics δ (such as the Spearman's footrule, the Spearman's rank or the Kendall's τ). The metric must satisfy the usual axioms (D1), (D2), (D3) and the right-invariance (D5) to build a φ-model.

The model belongs to an exponential family of probability distributions over $\mathfrak{S}_n$ and is unimodal. It can be seen as analogous to the normal distribution since the probability mass allocated to each permutation decreases exponentially with the distance to a central permutation π according to a *dispersion factor* φ. For the Kendall's τ, the probability of a total ranking σ in $\mathcal{M}(\pi, \varphi)$ writes as

$$\mathcal{M}(\pi, \varphi)(\sigma) := p_{\mathrm{M}}(\sigma | \pi, \varphi) := \frac{1}{Z} \varphi^{\tau_K(\pi, \sigma)}.$$

With Z being the normalization constant, which is the sum over $n!$ elements $\sum_{\sigma\in\mathfrak{S}_n}\varphi^{\tau_K(\pi,\sigma)}$. In the case of the Kendall's τ, the constant Z is explicitly known $Z=(1+\varphi)(1+\varphi+\varphi^2)\cdots(1+\varphi+\cdots+\varphi^{n-1})$. Note that similarly as for the normal distribution, the normalizing constant only depends on the scaling parameter φ. This is a natural property of a distance-based model for which only the distance to the location parameter π matters and not the parameter itself.

The extreme cases of the model are a uniform distribution over $\mathfrak{S}_n$ for $\varphi=1$ and a Dirac distribution δ_π when $\varphi\to 0$. Furthermore, the log-likelihood of the model is $\log\mathcal{L}(\pi)=\log\left(\frac{1}{Z}\varphi^{\tau_K(\pi,\sigma)}\right)=\tau_K(\pi,\sigma)\log\varphi-\log Z$, so for a profile of rankings $\mathcal{D}_L$,

$$\begin{aligned}F_{\text{MLE}}(\mathcal{D}_L)&=\operatorname*{arg\,min}_{\pi\in\mathfrak{S}_n}\sum_{l=1}^{L}\Big(\tau_K(\pi,\pi^{(l)})\log\varphi-\log Z\Big)\\&=\operatorname*{arg\,min}_{\pi\in\mathfrak{S}_n}\sum_{l=1}^{L}\tau_K(\pi,\pi^{(l)})\\&=\operatorname*{arg\,min}_{\pi\in\mathfrak{S}_n}d_K(\pi,\mathcal{D}_L).\end{aligned}$$

Therefore, the MLE estimator F_{MLE} of the mode of the Mallows model corresponds to a Kemeny order, *ie.* $F_{\text{MLE}}=F_{\text{KT}}$. Therefore, the CEMC algorithm is appropriate for recovering the central permutation of $\mathcal{M}(\pi,\varphi)$ in case the Kendall's τ distance is used.

The model is motivated by a generalization of the Condorcet model in which the preference pairs of a central permutation π are wrongly observed with *noise probability* $p\in\left[0,\frac{1}{2}\right[$. This model is said to be *noisy* since it assumes the existence of a ground truth of which the data are noisy observations. Mallows derived the φ-model in the Condorcet framework where the observations remain independent. Using that the observations of pairs are independent, we have the product over all pairs

$$\begin{aligned}p_{\text{M}}(\sigma|\pi,p)&=\tfrac{1}{Z'}\prod_{i\neq j\in[\![1,n]\!]}\begin{cases}p & \text{if } \pi \text{ and } \sigma \text{ disagree on the pair } (i,j)\\ 1-p & \text{otherwise}\end{cases}\\&=\tfrac{1}{Z'}p^{\tau_K(\pi,\sigma)}(1-p)^{\binom{n}{2}-\tau_K(\pi,\sigma)}\\&=\tfrac{1}{Z'}(1-p)^{\binom{n}{2}}\left(\tfrac{p}{1-p}\right)^{\tau_K(\pi,\sigma)}.\end{aligned}$$

Therefore, the dependency on σ is exponential with decay $\varphi:=\frac{p}{1-p}$. The relation between the normalization constants is $Z=Z'(1-p)^{-\binom{n}{2}}$. A naive way to sample permutations from a Mallows model $\mathcal{M}(\pi,\varphi)$ is to *noisily observe* all the pairs of π with the corresponding probability of error and eventually reject the whole observation if it is intransitive, *ie.* if the set of observed pairwise preferences doesn't constitute a ranking in $\mathcal{R}$. This procedure is highly expensive since exploring all the pairs costs $\mathcal{O}(n^2)$, testing if there is a cycle inside the set of pairwise preferences sampled is also costly and the rejection rate makes it unworkable.

Doignon et al. (2004, [18]) showed that the Repeated Insertion Model (RIM) allows to sample from $\mathcal{M}(\pi,\varphi)$ in a simpler way. We start from an empty ranking and incrementally insert $\pi_1,\pi_2,\cdots,\pi_n$ with probabilities $p_{i,j},\ i\in[\![1,n]\!],j\in[\![1,i]\!]$ depending on the pair (i,j). We need to ensure that the RIM corresponds to $\mathcal{M}(\pi,\varphi)$ by choosing the $p_{i,j}$ accordingly. Using the formalism of [19], we denote $\vec{j}=(j_1,\cdots,j_n)$ an *insertion vector* such as $\forall i\in[\![1,n]\!],j_i\leq i$ and $\varphi_\pi(\vec{j})$ is the total ranking obtained by successively inserting the π_i at the position given by j_i according to the RIM procedure. At each of these insertions, we create $i-j_i$ misorderings with respect to π, therefore we have $\tau_K(\pi,\varphi_\pi(\vec{j}))=\sum_{i=1}^{n}\left(i-j_i\right)$. Therefore we have a procedure based on RIM to efficiently sample from a Mallows model $\mathcal{M}(\pi,\varphi)$.

Algorithm 3 Repeated Insertion Model

```
procedure RIM($\pi, \varphi$)                                  ▷ samples from $\mathcal{M}(\pi, \varphi)$
    $\sigma \leftarrow \emptyset$
    for $i \in [\![1, n]\!]$ do
        insert $\pi_i$ into $\sigma$ at rank position $j$ with probability $\frac{\varphi^{i-j}}{1+\varphi+\cdots+\varphi^{i-1}}$ for $j \leq i$
    return $\sigma$
```

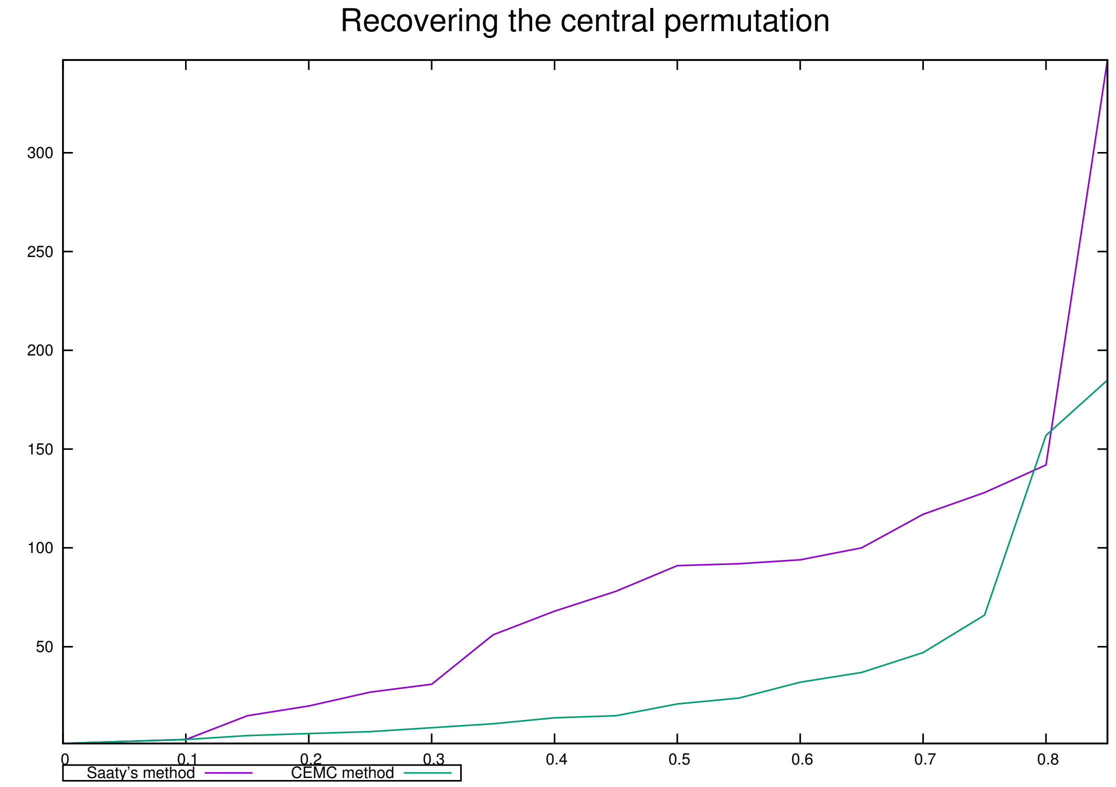


Figure 3: Number of noisy list observations L required to recover the central permutation π using the model $\mathcal{M}(\pi, \varphi)$ as a function of the dispersion factor φ going from a Dirac distribution for $\varphi = 0$ (a single permutation possible π) to the uniform distribution as φ goes to 1. Note that $\varphi \to 1$ is a non-informative case, hence the number L required soars. Comparison between Saaty's method (spectral method introduced later) and the CEMC method which is superior for data distributed as a Mallows model.

4 Utility-based aggregation

A *cardinal utility* function is a utilitarian score that preserves preference orderings up to positive affine transformations. It is used in theories of choice under risk for financial applications, notably for its ability to account for *how much* an alternative is better than an other while *ordinal utility* only accounts for *which* is better. The most famous application of cardinal utility is the *Expected Utility Theory* which offers the simplest way to quantify the risk aversion of an investor.

The problem of rank aggregation is formalized by the rank aggregation scheme $F : \mathcal{R}^L \to \mathcal{R}$ defined on rankings. Since the performance measures of strategies provide scores, *ie.* cardinal information rather than the ordinal information associated with rankings, we can directly aggregate these scores. We define the *score aggregation problem* as finding an aggregate scheme $A : \mathbb{R}^{n \times L} \to \mathbb{R}^n$ which takes L score vectors as input and produces an aggregated score vector.

The introduction of the Bradley-Terry model (1952, [20]) initiated the research on score-based aggregation. These methods are based on a pairwise data representing the preferences of individuals, initially used in the field of psychometrics. The preferences are encoded as a matrix J such that $J_{i,j}$ quantifies how much alternative i is preferred to alternative j. A score vector $s \in \mathbb{R}^n$ is then derived from J and the aggregated total ranking is produced by sorting the scores.

4.1 Least squares minimization

Mosteller (1951, [21]) proposed to encode the observed judgments into a skew-symmetric matrix $J = -J^\top$ and then to find a score vector s such that $\forall i, j \in [\![1, n]\!], s_i - s_j = J_{i,j}$. This method allows to cope with missing information by simply putting a *mask* $\Omega \in \{0, 1\}^{n \times n}$ on the matrix J, unknown elements $J_{i,j}$ are excluded from the optimization by null values of the mask using the Hadamard product $\otimes$. The score vector is computed by the least square minimization problem between the preference matrix J and the corresponding difference of scores, which can be written as a Frobenius norm

$$\begin{aligned} s &= \underset{x \in \mathbb{R}^n \ | \ x^\top \mathbf{1}}{\arg\min} \ \frac{1}{4} \sum_{i=1}^{n} \sum_{j=1}^{n} \Omega_{i,j} \Big(J_{i,j} - (x_i - x_j) \Big)^2 \\ &= \underset{x \in \mathbb{R}^n \ | \ x^\top \mathbf{1}}{\arg\min} \ \frac{1}{4} \| \Omega \ \otimes \ (J - \mathbf{1} x^\top - x \mathbf{1}^\top) \|_F^2 . \end{aligned}$$

Using the graph theory formulation, the mask is seen as an undirected graph whose vertices are the alternatives and the edges correspond to the presence of a preference judgment, *ie.* Ω is the *adjacency matrix*. The *degree matrix* of the graph is $D_\Omega := \operatorname{diag}(\Omega \mathbf{1})$, which allows us to formulate the gradient of the objective as $(D_\Omega - \Omega)x - (\Omega \ \otimes \ J)\mathbf{1}$. By definition, $D_\Omega - \Omega$ is a *Laplacian matrix* whose rows and columns sum to zero. Therefore setting $s := (D_\Omega - \Omega)^* (\Omega \ \otimes \ J)\mathbf{1}$ gives the solution to the minimization problem, since $\mathbf{1}^\top s = 0$.

4.2 Consistency of preferences

Since a profile of rankings $\mathcal{D}_L$ made of top-k lists gives us a complete matrix of pairwise information, we use a multiplicative analogous method due to Saaty (2003, [22]). The matrix of preferences $J \in \mathbb{R}_+^{n \times n}$ with positive components would ideally be of the form $\forall i, j, J_{i,j} = \frac{s_i}{s_j}$ where s is a preference vector.

The cardinal utility score s is a *priority vector* if it is invariant under the *hierarchic composition principle*, *ie.* if we weight the alternatives according to s and sum, the same score up to a positive constant must be produced. This hierarchic composition shouldn't produce new scores *ad infinitum*. Therefore s is such that $Js = cs$ with $c > 0$.

Saaty argued that under hierarchic composition invariance, there exists a unique priority vector. The theorem of Perron-Frobenius insures the unicity of the right eigenvector associated with the principal eigenvalue λ_{max}, called the *Perron root.* The main result of [22] is that the unique priority vector verifying $Js = cs$ is a positive multiple of the principal eigenvector and c is the Perron root λ_{max} itself. Note that the desired case $J_{i,j} = \frac{s_i}{s_j}$ directly gives $Js = ns$ and so n is the Perron root in this case.

The theorem is proved for the class of *consistent* matrices, *ie.* $\forall i, j, k \in [\![1, n]\!]$, $J_{i,j} = J_{i,k}J_{k,j}$ and is extended to the class of *near-consistent* matrices. A matrix J is near-consistent if it can be written as an Hadamard product $J = S \odot E$ with $E = (\epsilon_{i,j})$ a reciprocal multiplication perturbation matrix. Therefore we have $\forall i, j \in [\![1, n]\!]$, $J_{i,j} = \frac{s_i}{s_j}\epsilon_{i,j}$. First we notice that for all $i \in [\![1, n]\!]$,

$$\sum_{j=1}^{n} \epsilon_{i,j} = \sum_{j=1}^{n} J_{i,j}\frac{s_j}{s_i} = \frac{(Js)_i}{s_i} = \lambda_{max}.$$

So we have the following inequality

$$n\lambda_{max} = \sum_{i=1}^{n}\sum_{j=1}^{n} \epsilon_{i,j} = \sum_{i=1}^{n} \epsilon_{i,i} + \sum_{i<j} (\underbrace{\epsilon_{i,j} + \epsilon_{j,i}}_{= \, \epsilon_{i,j} + \epsilon_{i,j}^{-1} \, \geq \, 2}) \geq n + 2\frac{n(n-1)}{2} = n^2$$

which insures the bound $\lambda_{max} \geq n$. This motivates the definition of a *consistency index* $CI := \frac{\lambda_{max} - n}{n-1}$. Note that $CI = 0$ if and only if J is consistent.

To insure consistency, we assign positive judgment values into a reciprocal matrix $J_{i,j} = J_{j,i}^{-1}$ using an estimator of the odds ratio defined by Luce's Choice axiom, the *smoothed empirical ratios* writes as

$$J_{i,j} := \frac{\hat{p}_{i \succ j} + c}{\hat{p}_{j \succ i} + c}$$

where $\hat{p}_{i \succ j}$ denotes the empirical probability of $i \succ j$. The parameter c allows smoothing since the empirical ratios are levelled off towards 1 when $c \to +\infty$. The class of aggregation schemes based on J only requires the estimation of the pairwise marginals $p_{i \succ j}$ and therefore provides $n(n-1)/2$ accessible parameters. The marginals can naturally be estimated by a frequentist approach with the number of time the corresponding pairwise preference is observed in the input data $\mathcal{D}_L$.

The main right eigenvector of J can be computed as follows $s = \lim_{k \to \infty} \frac{J^k \mathbb{1}}{\mathbb{1}^\top J^k \mathbb{1}} \in \mathbb{R}^n$. We can then break the ties in the score vector by any permutation to recover an aggregated total ranking $F_{\mathrm{S}}(\mathcal{D}_L) \in \mathfrak{S}_n$. This procedure is a convenient score-based aggregation with low complexity $\mathcal{O}(n^2 \log n)$ but it only leverages pairwise information. Subsequent work focused on listwise models, estimating the probability of entire rankings, such as Random Utility Models.

4.3 Plackett-Luce model

For a given score $s \in \mathbb{R}^n$, the Plackett-Luce model is an extension of the multinomial logistic regression. It derives from the Luce's Choice axiom and writes as

$$p_{\mathrm{PL}}(\pi \mid s) := \prod_{i=1}^{n} \frac{s_{\pi_i}}{\sum_{j=i}^{n} s_{\pi_j}}.$$

A natural interpretation of the probability of the total ranking π is to pick the alternatives from top to bottom according to their score and by removing the alternative previously chosen at each step. By assigning a score to each alternative, the model has a global parameter $s \in \mathbb{R}^n$, whose components are estimated through top-k marginals. This approach differs from the distance-based Mallows φ-models whose parameters define a structure on the symmetric group $\mathfrak{S}_n$.

We first focus on the problem of estimating top-1 marginals from a top-1 partial ranking dataset $\mathcal{D}_L := \{(i^{(l)}, \mathcal{N}^{(l)}) \mid l \in [\![1, L]\!]\}$ where $i^{(l)}$ designates the best alternative of subset $\mathcal{N}^{(l)}$. From this partial data, Maystre *et al.* (2015, [23]) proposed to learn a Plackett-Luce model with a spectral algorithm, *ie.* leveraging on the pairwise preferences offered by the dataset. We maximize the log-likelihood of parameter $s \in \mathbb{R}^n$ given the dataset

$$\log \mathcal{L}(s \mid \mathcal{D}_L) := \log \prod_{l=1}^{L} \frac{s_{i^{(l)}}}{\sum_{i' \in \mathcal{N}^{(l)}} s_{i'}} = \sum_{l=1}^{L} \Big(\log s_{i^{(l)}} - \log \sum_{i' \in \mathcal{N}^{(l)}} s_{i'} \Big).$$

By taking the gradient with respect to s and rearranging the terms to sum over the pairs of alternatives (i, j), we can show the equivalence of the optimality condition $\log \mathcal{L}(s \mid \mathcal{D}_L) = 0$ and

$$\forall i \in [\![1, n]\!], \sum_{j \neq i} \Big(\sum_{l \in W_i \cap L_j} \frac{s_j}{\sum_{i' \in \mathcal{N}^{(l)}} s_{i'}} - \sum_{l \in W_j \cap L_i} \frac{s_i}{\sum_{i' \in \mathcal{N}^{(l)}} s_{i'}} \Big)$$

where $W_i := \{l \mid i^{(l)} = i, i \in \mathcal{N}^{(l)}\}$ and $L_i := \{l \mid i^{(l)} \neq i, i \in \mathcal{N}^{(l)}\}$ are the voters for which i wins over and lose against against the alternatives, respectively. We denote σ_S the weight function of s associated to a subset of the data $S \subseteq \mathcal{D}$ defined by

$$\sigma_S(s) := \sum_{\mathcal{N} \in S} \frac{1}{\sum_{i' \in \mathcal{N}} s_{i'}}.$$

We consider the batches of data associated to the pairwise marginals $\mathcal{D}_{i \succ j} := \{(i^{(l)}, \mathcal{N}^{(l)}) \in \mathcal{D}_L \mid l \in W_i \cap L_j\}$ to rewrite the optimality condition with the weight function. Let $J_{i,j}(s) := \sigma_{\mathcal{D}_{i \succ j}}(s)$ the judgment strength associated to the marginal $i \succ j$ for the score parameter s. The optimality condition can be written as the global balance conditions of the Markov chain associated with the inhomogeneous transition rates $J_{i,j}(s)$,

$$\forall i \in [\![1, n]\!], \sum_{j \neq i} s_i J_{j,i}(s) = \sum_{j \neq i} s_j J_{i,j}(s).$$

Ford (1957, [24]) studied the maximization of $\log \mathcal{L}(\theta; \mathcal{D}_L)$ and proposed the following necessary and sufficient condition on the data for the set of global optima to be bounded and unique - providing that one parameter is fixed, *eg.* $s_1 = 0$.

Ford's condition. For the top-1 partial ranking dataset, for any partition of the set of subsets $\{\mathcal{N}^{(l)}\}_{l=1}^{L} = \mathcal{X}_1 \sqcup \mathcal{X}_2$ into two non-empty sets, we have $\big(\cup_{\mathcal{N} \in \mathcal{X}_1} \mathcal{N} \big) \cap \big(\cup_{\mathcal{N} \in \mathcal{X}_2} \mathcal{N} \big) \neq \emptyset$.

This can be seen as requiring the hypergraph $([\![1, n]\!], \{\mathcal{N}^{(l)}\}_{l=1}^{L})$ to be connected. Maystre *et al.* proved that under Ford's condition and the marginal distributional assumption, the stationary distribution $\bar{s}$ of the time-inhomogeneous Markov chain associated to the preference matrix $J(s)$ coincides with the maximum likelihood score, *ie.* $\bar{s} = s^*$. The Luce Spectral Ranking algorithm relies on this property.

Algorithm 4 Luce Spectral Ranking

```
procedure LSR($\mathcal{D}_L$)
    $J \leftarrow 0 \in \mathbb{R}^{n \times n}$
    for $(i, \mathcal{N}) \in \mathcal{D}_L$ do
        for $j \in \mathcal{N} \setminus \{i\}$ do
            $J_{i,j} \leftarrow J_{i,j} + n/|\mathcal{N}|$        ▷ build preference matrix
    $\bar{s} \leftarrow$ stationary distribution of Markov chain $J$
    return $\bar{s}$
```

The stationary distribution of the Markov chain can be computed using the eigenvalue method or a LU decomposition. The main idea of this aggregation scheme is to build preferences with respect to a structural model depending on parameter s and then *calibrating* s to respect the invariant composition principle. Therefore, the rates must be updated at each iteration to take into account the changing preferences $J(s)$. The main limitation of this algorithm lies in the use of pairwise information. In order to fully exploit the ranking information of a dataset, we propose to estimate the probability of an entire ranking with the help of the Random Utility theory.

4.4 Random Utility Models

Random utility theory has been developed in economics to overcome the complexity limitations of the Kemeny and Condorcet models. Random Utility Models (RUMs) construct an agent's preferences by drawing scores from a distribution parameterized by a ground truth. The use of RUMs received a decisive momentum from the breakthrough contribution of Soufiani *et al.* (2012, [25]) who devised an efficient EM procedure to learn a wide class of multivariate distributions from the exponential family, including the Plackett-Luce model.

Random Utility Models are defined from a ground truth vector $\theta \in \mathbb{R}^n$ used to compute the probability of a total ranking, *ie.* an n^{th} order probability marginal. The estimation of listwise probabilities allows to use more ranking information and therefore extracts more information from low consistency datasets, *ie.* such that $CI > 0$.

The probability of producing a given total ranking π with the random utilities $(X_{\pi_1} > \cdots > X_{\pi_n})$ is

$$\begin{aligned} p_{\text{RUM}}(\pi \mid \theta) &= \mathbb{P}(X_{\pi_1} > \cdots > X_{\pi_n}) \\ &= \int_{x_{\pi_n}=-\infty}^{\infty} \int_{x_{\pi_n}=x_{\pi_{n-1}}}^{\infty} \cdots \int_{x_{\pi_1}=x_{\pi_2}}^{\infty} \mu_{\pi_n}(x_{\pi_n}) \cdots \mu_{\pi_1}(x_{\pi_1}) \, \mathrm{d}x_{\pi_1} \cdots \mathrm{d}x_{\pi_n} \end{aligned}$$

where $\mu_i = \mu(\cdot \mid \theta_i)$ is the parameterized distribution for random utility X_i using θ_i as a mean.

In the usual MLE approach to social choice, we learn the parameter θ which maximizes the likelihood of the observed profile of rankings $\mathcal{D}_L$, $p_{\text{RUM}}(\mathcal{D}_L \mid \theta) = \prod_{l=1}^{L} p_{\text{RUM}}(\pi(l) \mid \theta)$. Following the work of Soufiani *et al.* we consider distributions μ_i belonging to the exponential family

$$\mu_i(x) = \exp(\eta(\theta_i)T(x) - A(\theta 8i) + B(x)).$$

The Plackett-Luce model can be cast as a RUM with Gumbel distributions $\mu_i(x) = e^{-(x-s_i)} \exp(-e^{-(x-s_i)})$ and the corresponding part functions are $\eta(\theta_i) = s_i = e^{\theta_i}$, $T(x_i) = -e^{-x_i}$, $B(x_i) = -x_i$ and $A(\theta_i) = -\theta_i$. The Gumbel distribution is used to model errors in discrete choice theory because the difference of two Gumbel distributed random variables follows a logistic distribution.

4.5 Parameter estimation of RUMs

We restrict again the class of distributions to the *location family* where the scale parameters of the distributions are fixed. The i^{th} random utility can be written as $X_i = \theta_i + \xi_i$ where ξ_i is a centered *subjective noise*. To design an efficient learning procedure, we need the concavity of the log-likelihood $\log \mathcal{L}(\theta \mid \mathcal{D}_L) = \sum_{l=1}^{L} \log p_{\text{RUM}}(\pi^{(l)} \mid \theta)$. It is sufficient to have each ξ_i distributed with a log-concave density. The Gumbel and fixed-shape normal $\{\mathcal{N}(\theta, \sigma^2) \mid \theta \in \mathbb{R}\}$ families belong to the location family. Ford's necessary and sufficient condition for the set of global maxima solutions to be bounded in the case of the Gumbel distributions writes as follows.

Ford's condition. Given the data $\mathcal{D}_L$, for any partition of $[\![1, n]\!] = \mathcal{X}_1 \sqcup \mathcal{X}_2$ of two non-empty subsets of alternatives, there exists $i \in \mathcal{X}_1, j \in \mathcal{X}_2$ such that there is at least one total ranking $\pi^{(l)} \in \mathcal{D}_L$ where $i \succ_{\pi^{(l)}} j$.

Estimating the parameter θ can be achieved with an algorithm which builds a sequence $\{\theta^{(t)}\}_{t=1}^{\infty}$ by iterating the following EM procedure.

- E-STEP. For any $\theta \in \mathbb{R}^n$, we compute the conditional expectation of the log-likelihood of the augmented data ($\mathcal{D}_L$ and the latent variables X). The latent variables X are distributed conditionally on $\mathcal{D}_L$ and the current selected parameter $\theta^{(t)}$

$$Q(\theta, \theta^{(t)}) = \mathbb{E}_X \Big[\log \prod_{l=1}^{L} p(X^{(l)}, \pi^{(l)} \mid \theta) \Bigm| \mathcal{D}_L, \theta^{(t)} \Big].$$

- M-STEP. We solve the optimization problem associated to $Q(\cdot, \theta^{(t)})$ and use the result for the next selected parameter $\theta^{(t+1)}$

$$\theta^{(t+1)} \in \operatorname*{arg\,max}_{\theta \in \mathbb{R}^n} Q(\theta, \theta^{(t)}).$$

Using that μ belongs to the exponential family, the conditional expectation can be decomposed as

$$\begin{aligned} Q(\theta, \theta^{(t)}) &= \mathbb{E}_X \Big[\log \prod_{l=1}^{L} p(X^{(l)} \mid \theta) p(\pi^{(l)} \mid X^l) \Bigm| \mathcal{D}_L, \theta^{(t)} \Big] \\ &= \sum_{l=1}^{L} \sum_{i=1}^{n} \mathbb{E}_{X_i^l} \Big[\log \mu_i(X_i^{(l)}) \Bigm| \pi^{(l)}, \theta^{(t)} \Big] \\ &= \sum_{l=1}^{L} \sum_{i=1}^{n} \Big(\eta(\theta_i) \mathbb{E}_{X_i^l} \Big[T(X_i^{(l)}) \Bigm| \pi^{(l)}, \theta^{(t)} \Big] - A(\theta_i) + C \Big) \end{aligned}$$

where $C := \mathbb{E}_{X_i^{(l)}} \Big[B(X_i^{(l)}) \Bigm| \pi^l, \theta^{(t)} \Big]$ is independent of θ and doesn't intervene in the optimization.

It remains to estimate the conditional expectations $S_i^{(l),(t+1)} := \mathbb{E}_{X_i^{(l)}} \Big[T(X_i^{(l)}) \Bigm| \pi^{(l)}, \theta^{(t)} \Big]$ where T is the sufficient statistic bearing the information of the shape of μ_i. Using a Monte Carlo estimation, we have

$$S_i^{(l),(t+1)} \approx \frac{1}{N} \sum_{k=1}^{N} T(x_i^{(l),k}) \text{ with } \{x_i^{(l),k}\}_{k=1}^{N} \sim p_{\text{RUM}}(\cdot \mid \pi^{(l)}, \theta^{(t)}).$$

Soufiani *et al.* proposed several technical details to carry the estimation of the $S_i^{(l),(t+1)}$ efficiently. First, the sampling from the conditional distribution $p(\cdot \mid \pi^{(l)}, \theta^{(t)})$ is achieved with a Gibbs sampler. For each $k \in [\![1, N]\!]$, we select an alternative i in $\pi^{(l)}$ then sample its utility $x^{(l)}_{\pi_i^l}$ from a truncated version of $\mu_{\pi_i^{(l)}}(\cdot)$ at the previous and next utilities $x^{(l)}_{\pi_{i-1}^{(l)}}$ and $x^{(l)}_{\pi_{i+1}^{(l)}}$. We denote the distribution obtained at step k of the Gibbs sampler by $p_{\text{RUM,tr}}(\cdot \mid x_{-i}^{l(l),k}, \pi^l, \theta^{(t)})$ where $x_{-i}^{(l),k}$ is the current vector of sampled utilities without the utility of alternative i.

4.6 Normal RUM with fixed variance

We consider the case of the normal distribution with fixed variance σ^2 which belongs to the exponential family since

$$\mathcal{N}(\theta, \sigma^2)(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\theta)^2}{2\sigma^2}} \implies \eta(\theta) = \frac{\theta}{\sigma^2},\ T(x) = x,\ A(\theta) = \frac{\theta^2}{2\sigma^2},\ B(x) = \frac{-x^2}{2\sigma^2} - \frac{1}{2}\log(2\pi).$$

The normalization constant only accounts in the *base measure* $B(x)$ which doesn't intervene in the E-step equations. The truncated normal distribution can easily be sampled using a rejection-based sampling of the normal distribution $\mu_{\pi_i^{(l)}}$. The E-step consists into estimating the $S_i^{(l),(t+1)}$ and the M-step consists into solving the optimization problem

$$\theta^{(t+1)} \in \operatorname*{arg\,max}_{\theta \in \mathbb{R}^n} \sum_{l=1}^{L} \sum_{i=1}^{n} \Big(\eta(\theta_i) S_i^{(l),(t+1)} - A(\theta_i) \Big).$$

In the case of the family $\{\mathcal{N}(\theta, \sigma^2) \mid \theta \in \mathbb{R}\}$, the estimator of the ground truth parameter θ used as next iteration parameter $\theta^{(t+1)}$ is given by

$$\forall i \in [\![1, n]\!],\ \theta_i^{(t+1)} = \frac{1}{L} \sum_{l=1}^{L} S_i^{(l),(t+1)}.$$

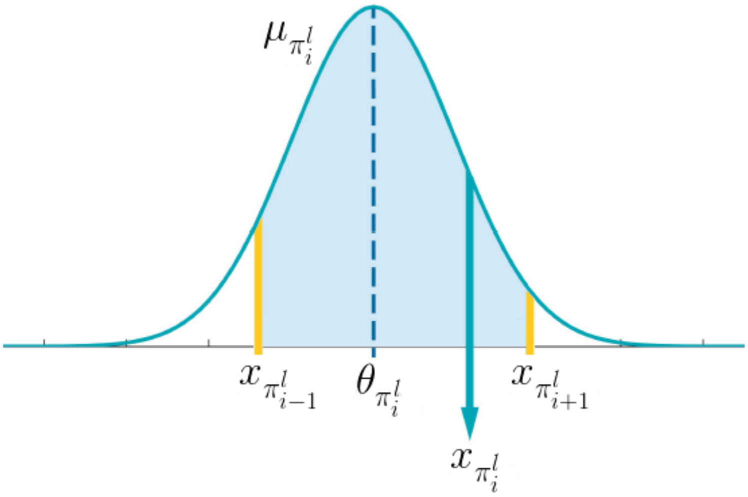


Figure 4: Example of density of a truncated normal distribution $p_{\text{RUM,tr}}(\cdot \mid x_{-i}^{(l),k}, \pi^{(l)}, \theta)$.

The procedure derived to learn the parameter θ from a profile of rankings $\mathcal{D}_L$ using a Random Utility Model belongs to the class of *Minorize-Maximization* (MM) algorithms since at each iteration t, the conditional expectation $Q(\cdot, \theta^{(t)})$ of the E-step is a surrogate function which minorizes the log-likelihood objective. Once the parameter θ is learned, an aggregate ranking is obtained by simply sorting its components. This algorithm allows an efficient calibration of the normal RUM and is one of the leading utility-based rank aggregation schemes.

Algorithm 5 Monte Carlo EM algorithm for learning normal RUM

```
procedure MC-EM($\epsilon, N$)
    $t \leftarrow 1, \theta^{(1)} \leftarrow 0$
    while $\|\theta^{(t+1)} - \theta^{(t)}\| > \epsilon$ do                          ▷ Stopping criterion
        for $l \in [\![1, L]\!]$ do                              ▷ E-step, parallelizable Gibbs sampler
            for $k \in [\![1, N]\!]$ do
                $i \sim \mathcal{U}_{[\![1,n]\!]}$                       ▷ sample an alternative uniformly
                $x_i^{(l),k} \sim p_{tr}(\cdot \mid x_{-i}^{(l),k}, \pi^{(l)}, \theta^{(t)})$
            for $i \in [\![1, n]\!]$ do
                $S_i^{(l),(t+1)} \leftarrow \frac{1}{N} \sum_{k=1}^{N} T(x_i^{(l),k})$     ▷ Monte Carlo approximation
        for $i \in [\![1, n]\!]$ do                                                ▷ M-step
            $\theta_i^{(t+1)} \leftarrow \frac{1}{L} \sum_{l=1}^{L} S_i^{(l),(t+1)}$
    $T \leftarrow t$
    return $\theta^{(T)}$
```

5 Learning to rank formulation

The proposed solutions to the ranking aggregation problem allow to judge the quality of strategies with respect to one another. The aggregation schemes must be called each time new strategies are to be ranked and are expensive in the number of strategies. Using *online learning* algorithms, a model can be learned beforehand to rank a set of strategies in constant time. The algorithms require to select features describing the alternatives. In the case of portfolio optimization, these features must describe the distribution of the expected returns of an investment strategy. Furthermore, we would like to leverage on the possibility to assess every portfolio with each performance measure, a significant difference from the field of Information Retrieval where very few preferences are provided by the users.

The problem of ordering a set of alternatives based on preferences is ubiquitous in the fields of application of Machine Learning, nonetheless our comprehension of learning to rank falls short of our comprehension of the usual regression and classification problems. The problem is often tackled with pointwise formulations, such as casting it into a classification problem (*is this alternative relevant or not?*) or with the ordinal regression.

Tie-Yan Liu (2008, [26]) categorized the learning to rank algorithms into three groups depending on their cost function.

- Pointwise approach: each datum in the training set is associated with a score and the learning problem boils down to a regression - predict the score of a new single alternative.
- Pairwise approach: the learning problem boils down to a classification problem - a classifier is learned to predict which alternative is better when a pair of alternative is presented. It corresponds to minimizing the number of inversions in the final ranking.
- Listwise approach: the learning problem minimizes a cost metric defined on lists. This is the most natural way to tackle the problem of learning to rank but also the most challenging. Most metrics being highly discontinuous, surrogate losses must be considered.

We propose to adapt the *Learning to Rank* paradigm to the portfolio optimization problem by focusing on listwise loss functions. Several studies have suggested that optimization with respect to listwise losses produces better result than with respect to pairwise or pointwise losses. A theoretical approach is suggested to derive an optimal algorithm insuring consistency of the surrogate loss function. Capturing information of full ranks is computationally intractable but we can use approximations of top-k probabilities to capture partial rankings information.

5.1 Feature space and dataset

As in the work of Duchi (2013, [1]), we consider that the user provides a query q conditionally on which the alternatives must be ranked. We assume a finite number of queries so they can be labeled by a *query index* $q \in [\![1, Q]\!]$. We then work in a *features space* $\mathcal{X}$ of finite dimension d with each alternative $i \in [\![1, n]\!]$ represented by a vector of features $x_i^{(q)} \in \mathcal{X}$ conditional on the query q. The features are chosen to represent an alternative with a finite set of characteristics. For a query q, $\mathcal{N}^{(q)}$ denotes the subset of the alternatives corresponding to q.

We are looking to learn a mapping $h : q \in [\![1, Q]\!] \mapsto \mathfrak{S}_{n^{(q)}}$ that would produce the desired ranking of the subset of alternatives $\mathcal{N}^{(q)}$. This mapping can be learned from a query-based dataset providing a batch of data $\{x_i^{(q)}, y_i^{(q)}\}_{i=1}^{n^{(q)}}$ for each query q, where $y_i^{(q)}$ is a *relevance score* associated to the alternative $x_i^{(q)}$. The usual $0-1$ loss function for rankings $l(h(q), y) := \mathbb{I}(h(q) \neq y)$ is defined by comparing the permutation produced $h(q)$ with the permutation associated to the vector of scores $y^{(q)}$. It is the listwise loss used for the theoretical analysis of Learning to Rank.

Therefore, we optimize the *empirical risk* associated with the mapping h

$$R(h) = \frac{1}{Q}\sum_{q=1}^{Q} l(h(q), y^{(q)}).$$

As in the work of Xia *et al.* (2008, [27]), for efficiency consideration we consider that the ranking function h is decomposable with respect to alternatives. We define a scoring function $s : \mathbb{R}^d \to \mathbb{R}$ and use it to score each alternative into a score vector $s(x^{(q)}) := (s(x_1^{(q)}), \cdots, s(x_{n^{(q)}}^{(q)}))$. In the population formulation, this vector is chosen from a vector Borel-measurable function set.

We can extract a permutation from this score vector by sorting its components to produce the mapping $h_s(q) := \mathrm{sort}(s(x^{(q)}))$. This class of mappings naturally arises from the utility-based theories studied previously. We proceed to optimize the associated risk with respect to the scoring function s instead

$$R(s) = \frac{1}{Q}\sum_{q=1}^{Q} l(h_s(q), y^{(q)}) \text{ with } h_s(q) := \mathrm{sort}(s(x^{(q)})).$$

5.2 Listwise surrogate losses

The $0-1$ loss makes this risk $s \mapsto R(s)$ non-differentiable and therefore its minimization intractable. We replace the loss function l by a surrogate loss φ between the ranking ground truth seen as a permutation $y^{(q)} \in \mathfrak{S}_{n^{(q)}}$ and the score vector produced $s(x^{(q)})$ which gives the surrogate risk

$$R^{\varphi}(s) = \frac{1}{Q}\sum_{q=1}^{Q} \varphi^{(q)}(s(x^{(q)}), y^{(q)})$$

where $\varphi^{(q)} : \mathbb{R}^{n^{(q)}} \times \mathfrak{S}_{n^{(q)}} \to \mathbb{R}$ is the surrogate loss. Note that it depends on the query q since the size of the subset of alternatives varies. In the following sections, the subscript (q) is often left out to alleviate the notations but the analysis is carried out for any query q on the subset of alternatives $\mathcal{N}^{(q)}$ relabeled as $[\![1, n^{(q)}]\!]$.

The surrogate loss φ is chosen to ensure tractability of the optimization and Fisher-consistency, *ie.* solving the surrogate problem yields a solution to the initial risk minimization problem. This is formalized by the following definition from [28]. A surrogate loss φ is *Fisher-consistent* with respect to the $0-1$ loss if

$$s^* \in \underset{s\,:\,\mathbb{R}\to\mathbb{R}}{\arg\min}\ \mathbb{E}_{q,Y\sim p}\Big[\varphi(s(q), Y)\Big] \implies s^* \in \underset{s\,:\,\mathbb{R}\to\mathbb{R}}{\arg\min}\ \mathbb{E}_{q,Y\sim p}\Big[l(h_s(q), Y)\Big]$$

Xia *et al.* gave sufficient conditions for φ and the probability space to be Fisher-Consistent with respect to the $0-1$ loss. For alternatives i, j, we define the marginal set of permutations $\mathfrak{S}_{i\succ j} := \{y \in \mathfrak{S}_n \mid i \succ j\}$ corresponding to the condition $i \succ j$. The probability space $(\Omega, \mathcal{F}, p)$ is *order-sensitive* if

$$\forall i, j \in \mathcal{N}, \forall y \in \mathfrak{S}_{i\succ j},\ p(y) > p(\sigma_{i,j} y)$$

where $\sigma_{i,j} y$ denotes the permutation obtained by swapping alternatives i and j in y and keeping the others unchanged. This means that for each pair i, j, the probability measure must favor the permutations ranking the pair in the *right order* defined by $\succ$ over the same permutations with the exception of the swapped pair.

Furthermore, a surrogate loss is *order-sensitive* if

$$\begin{cases} \forall i \succ j,\ \forall y \in \mathfrak{S}_n,\ s(x_{y_i}) < s(x_{y_j}) \implies \varphi(s, y) \geq \varphi(s, \sigma_{i,j} y) \\ \forall i, j,\ s_i = s_j \implies \begin{cases} \forall y \in \mathfrak{S}_{i\succ j}, \frac{\partial}{\partial s_i}\varphi(s, y) \leq \frac{\partial}{\partial s_j}\varphi(s, y) \\ \text{or } \forall y \in \mathfrak{S}_{i\succ j}, \frac{\partial}{\partial s_i}\varphi(s, y) \geq \frac{\partial}{\partial s_j}\varphi(s, y) \end{cases} \end{cases}$$

where all conditions must be strict for at least one y. This means that if alternative y_i is better than alternative y_j according to the permutation y and if the scores $s(x_{y_i})$ and $s(x_{y_j})$ disagree, we can decrease the loss φ by swapping i and j in y.

Xia's consistency theorem. If both the probability space and the surrogate loss are order-sensitive then the surrogate loss is Fisher-consistent with respect to the $0-1$ loss ([27]).

Using log-likelihood, we can build consistent losses $\varphi(s(x), y) = -\log p(y|x, s)$ using order-sensitive probabilistic models p. The Plackett-Luce model $(\Omega, \mathcal{F}, p_{\text{PL}})$ is order-sensitive *ie.* if $i \succ j$ and $s(x_{y_i}) < s(x_{y_j})$, since the score attributed to $x_{\sigma_{i,j}y_i}$ is larger than the score of $x_{\sigma_{i,j}y_j}$, the modified permutation $\sigma_{i,j}y$ is more likely. The model has other desirable properties, such as $y = (y_1, \cdots, y_n)$ is the most likely permutation and $(y_n, \cdots, y_1)$ the most unlikely. With an exponential score, the model can be written using the softmax function

$$p_{\text{PL}}(y \mid x, s) = \prod_{i=1}^{n} \sigma(s_{x_y})_i \text{ with } \sigma(s)_i = \frac{e^{s_i}}{\sum_{j=1}^{n} e^{s_j}}.$$

5.3 Top-k probabilities

In order to simplify the optimization while retaining as much ranking information as possible, we work with top-k probabilities, *ie.* the probability of the set of permutations whose top-k elements are exactly $j_1, \cdots, j_k$. All such permutations form the subgroup $\mathfrak{S}(j_1, \cdots, j_k)$ whose probability in a ranking model p is the marginal $p(\mathfrak{S}(j_1, \cdots, j_k) \mid s) := \sum_{\pi \in \mathfrak{S}(j_1, \cdots, j_k)} p(\pi \mid s)$. Fortunately, this top-$k$ marginal probability is easily computable in the case of the Plackett-Luce model using the following property, due to Cao *et al.* (2007, [29])

$$p_{\text{PL}}(\mathfrak{S}(j_1, \cdots, j_k) \mid s) = \prod_{i=1}^{k} \frac{\exp(s(x_{j_i}))}{\sum_{i'=i}^{n} \exp(s(x_{j_{i'}}))}.$$

Note that $(j_{k+1}, \cdots, j_n)$ used in the expression only appear in the normalization constant and can be taken in any order to normalize in the softmax function. We define the Cross Entropy loss φ_{CE} between the top-k probability distributions with respect to a relevance score y and a predicted score $z := s(x)$ by

$$\varphi_{\text{CE}}(y, z) = -\sum_{g \in \mathfrak{S}_k} p_{\text{PL}}(g \mid y) \log(p_{\text{PL}}(g \mid z))$$

where $\mathfrak{S}_k$ denotes the collection of the top-k subgroups. Computing this loss costs $\mathcal{O}(\frac{n!}{(n-k)!} n)$ since the collection contains $\frac{n!}{(n-k)!}$ subgroups whose probabilities are computable in $\mathcal{O}(n)$ (each top-k subgroup contains $(n-k)!$ permutations).

The model ListNet learns a linear prediction function $z = s(x) := \langle \omega, x \rangle$ with $\omega \in \mathbb{R}^d$ by minimizing φ_{CE} with a gradient descent. In the general case of top-k probabilities, the gradient with respect to ω is written as

$$\nabla_\omega \varphi_{\text{CE}}(y, z) = -\sum_{g \in \mathfrak{S}_k} \Big(\prod_{i=1}^{k} \hat{\sigma}(y)_i\Big)\Big(\sum_{i=1}^{k} (x_{j_i} - \sum_{i'=i}^{n} \hat{\sigma}(z)_{i'} x_{j_{i'}})\Big)$$

where $\hat{\sigma}(s)$ denotes the *partial softmax vector* of the score vector s and is defined as $\hat{\sigma}(s)_i = \frac{e^{s_i}}{\sum_{j=i}^{n} e^{s_j}}$.

Computing this gradient $\nabla_\omega \varphi_{\text{CE}}(y, z)$ costs $\mathcal{O}(k! k n^2 p)$, therefore stochastic gradient descent is needed to cope with the factorial complexity of the top-k loss. The top-1 is often used to determine the most likely alternative when this choice suffices, but most of the ranking information is lost since learning the top-1 alternative doesn't leverage on partial rankings. Using that we sum over the n top-1 permutations

$g = (j_1, \cdots)$, we can derive a simple form for the top-1 gradient with a single sum, computable in $\mathcal{O}(n)$

$$\begin{aligned}
\nabla_\omega \varphi_{\text{CE}}(y, z) &= - \sum_{g \in \mathfrak{S}_k} \frac{\partial p_{\text{PL}}(g \mid z)}{\partial \omega} \frac{p_{\text{PL}}(g \mid y)}{p_{\text{PL}}(g \mid z)} \\
&= \sum_{g \in \mathfrak{S}_k} \frac{e^{y_{j_1}}}{\sum_{i=1}^{n^{(q)}} e^{y_{j_i}}} \Big(\frac{\sum_{i=1}^{n} e^{z_{j_i}} x_{j_i}^{(q)}}{\sum_{i=1}^{n} e^{z_{j_i}}} - x_{j_1} \Big) \\
&= \sum_{g \in \mathfrak{S}_k} \sigma(y)_{j_1} \Big(\sum_{i=1}^{n} \sigma(z)_{j_i} x_{j_i} - x_{j_1} \Big) \\
&= \sum_{j_1=1}^{n} \Big(\sigma(y)_{j_i} - \sigma(z)_{j_i} \Big) x_{j_1}.
\end{aligned}$$

Since learning the top alternative considerably reduces the learning capacity of the listwise framework, Luo *et al.* (2015, [30]) proposed to sample a few top-k subgroups to evaluate an approximation of the top-k gradient loss at a fixed cost, hence breaking the factorial complexity in k. We sample S top-k subgroups $g_1, \cdots, g_S \in \mathfrak{S}_k$ by sampling the top-k alternatives of each g. The probability distribution used for the sampling is derived from the current score learned z in what Luo *et al.* called the *Adaptive distribution sampling*. An alternative i whose predicted score z_i is high is more likely to be sampled among the top-k alternatives of the samples $(g_1, \cdots, g_S)$. This Monte Carlo procedure uses S samples, which doesn't depend on k. This induces stochasticity in the choice of the direction of descent so the procedure is a stochastic gradient descent which costs $\mathcal{O}(Skn^2p)$.

Algorithm 6 Stochastic Top-k ListNet

```
procedure ListNet($k, \epsilon, \eta$)
    $\omega \leftarrow \mathcal{N}(0, \epsilon)$                                   ▷ randomly initialize the model
    for $t \in [\![1, T]\!]$ do                                                    ▷ $T$ iterations
        for $q \in [\![1, Q]\!]$ do
            for $i \in [\![1, n^{(q)}]\!]$ do
                sample the $S$ permutation subgroups $g_1, \cdots, g_S \in \mathfrak{S}_k$
                compute the prediction $z^{(q)} = \langle \omega, x^{(q)} \rangle$      ▷ $x^{(q)}$ is the vector of features knowing $q$
                compute the gradient $\nabla \varphi_{\text{CE}} := \nabla_\omega \varphi_{\text{CE}}(y^{(q)}, z^{(q)})$
                update the model $\omega \leftarrow \omega - \eta \nabla \varphi_{\text{CE}}$
    return $\omega$                                                                ▷ return learned model
```

A simple Tychonoff regularization $\frac{\lambda}{2}\|\omega\|_2^2$ is often added to the model to avoid overfitting and computational instability in the softmax function (the exponential normalization is highly sensitive to overflows).

5.4 Normalized Discounted Cumulative Gain

The graded label vector y provides relevance score for each alternative. For a predicted score vector z, we denote $\pi_z(i)$ the rank of alternative i in the ordering induced by z. For a given ranking, we call *cumulative gain* at precision k the sum of the relevance scores of the top-k alternatives and denote it $CG_k(y) := \sum_{i=1}^{k} y_i$. We emphasize the relevance scores with a monotonically increasing function $G(r) := 2^r - 1$ and discount each contribution by its position $\pi_z(i)$ in z with another monotonically increasing function $F(i) := \log_2(1 + i)$. This gives us the *Discounted Cumulative Gain* at precision k of the ranking

$$DCG_k(z, y) := \sum_{i=1}^{k} \frac{G(y_i)}{F(\pi_z(i))} = \sum_{i=1}^{k} \frac{2^{y_i} - 1}{\log_2(1 + \pi_z(i))}.$$

For a query q, the DCG at full precision $k = n^{(q)}$ must be normalized with respect to the query to obtain the commonly used *Normalized Discounted Cumulative Gain* in our query-based setup

$$NDCG(z, y) := \frac{1}{Z^{(q)}} \sum_{i=1}^{n^{(q)}} \frac{2^{y_i} - 1}{\log_2(1 + \pi_z(i))}$$

where Z is the normalization constant, *ie.* the maximum DCG over all rankings of the alternatives $\mathcal{N}^{(q)}$. This normalization with F has first been used in the field of Information Retrieval and is crucial to achieve consistency. Ravikumar *et al.* (2011, [31]) showed that the choice of the discount function F is critical in the designing of consistent surrogate losses to optimize with respect to the NDCG metric. The functions F and G define a family of NDCG metrics: the precision at k can be taken into account directly by the discount function with $F(i) = +\infty$ if $i > k$ which cancels out the remaining alternatives. For example the usual Prec@k metric is member of the NDCG family with $G(r) = r$ and $F(i) = \mathbb{I}(i \leq k)$.

We propose to minimize the NDCG at full precision (all alternatives considered for each query) seen as a loss, which corresponds to maximizing the gain. We denote this loss by $\varphi_{\text{NDCG}}(z, y)$ where $\pi_z^{(q)}(i)$ is the rank of alternative i for query q (this is tantamount to summing from top to bottom along the ranking induced by z).

$$\varphi_{\text{NDCG}}(z, y) := 1 - \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{Z^{(q)}} \sum_{i=1}^{n^{(q)}} \frac{2^{y_i^{(q)}} - 1}{\log_2(1 + \pi_z^{(q)}(i))}$$

Since the loss depends on the ranking induced by the score vector z and not directly on the score, the optimization is challenging. Valizadegan *et al.* (2009, [32]) proposed a probabilistic framework to optimize with respect to this criterion by considering the expectation over all the possible rankings for a given query q. We assume the existence of a ranking model $p^{(q)}(\cdot \mid s)$ knowing the scoring function s and the query q.

The expected loss can be written using this ranking model

$$\begin{aligned}
\mathbb{E}_z\Big[\varphi_{\text{NDCG}}(z, y)\Big] &= 1 - \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{Z(y^{(q)})} \sum_{i=1}^{n^{(q)}} \mathbb{E}\left[\frac{2^{y_i^{(q)}} - 1}{\log_2(1 + \pi_z^{(q)}(i))}\right] \\
&= 1 - \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{Z(y^{(q)})} \sum_{i=1}^{n^{(q)}} \left(\sum_{\pi \in \mathfrak{S}_{n^{(q)}}} p(\pi \mid s, q) \; \frac{2^{y_i^{(q)}} - 1}{\log_2(1 + \pi_i)} \right).
\end{aligned}$$

The optimization is based on the following bound which holds for all distributions p, using Jensen's inequality

$$\mathbb{E}_z\Big[\varphi_{\text{NDCG}}(z, y)\Big] \leq 1 - \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{Z(y^{(q)})} \sum_{i=1}^{n^{(q)}} \frac{2^{y_i^{(q)}} - 1}{\log_2(1 + \mathbb{E}[\pi_i])}.$$

Then by writing the rank as $\pi_i = 1 + \sum_{j=1}^{n} \mathbb{I}\big(\pi_i > \pi_j\big)$, we can express the expectation of the position π_i using the pairwise marginals of the probability model $p(j \succ_\pi i)$. This approach is based on a probability model for which the pairwise ranking probabilities are directly dependent on the function h predicting the score z

$$\mathbb{E}[\pi_i] = 1 + \sum_{j=1}^{n} \mathbb{E}\Big[\mathbb{I}\big(\pi_i > \pi_j\big)\Big] = 1 + \sum_{j=1}^{n^{(q)}} p(j \succ_\pi i).$$

We denote by S_λ the sigmoid function with shape parameter λ such as $S_\lambda(\delta) := \frac{1}{1+e^{-\lambda\delta}}$. We can consider a framework which easily models the pairwise marginals, for example by considering them independent with the following model

$$\forall q \in [\![1, Q]\!],\; p^{(q)}(\pi \mid s) = \frac{1}{Z(s, q)} \; \exp\left(\sum_{i=1}^{n^{(q)}} \sum_{\substack{j=1 \\ \pi_j > \pi_i}}^{n^{(q)}} \big(s(x_i^{(q)}) - s(x_j^{(q)})\big) \right).$$

In this case, assuming $s(x_i^{(q)}) > s(x_j^{(q)})$, the corresponding pairwise marginal satisfies the inequality

$$p(j \succ_\pi i) \leq S_2(s(x_j^{(q)}) - s(x_i^{(q)})).$$

It motivates the approximation with the logistic function of the left side of the inequality. [32] gives the theoretical justification of this logistic model widely used to approximate such marginals in ranking statistics. By plugging this approximation into the previous inequality, we have to minimize the upper bound

$$H(s) := 1 - \frac{1}{Q}\sum_{q=1}^{Q} \frac{1}{Z(y^{(q)})} \sum_{i=1}^{n^{(q)}} \frac{2^{y_i^{(q)}} - 1}{\log_2(1 + A_i^{(q)}(s))}$$

where $A_i^{(q)}(s) := \sum_{j=1}^{n^{(q)}} S_2(s(x_j^{(q)}) - s(x_i^{(q)}))$ is the approximation of the expectation $\mathbb{E}[\pi_i]$. Using a Taylor approximation of the discount factor, the problem consists of minimizing a quantity $M(s)$ not depending on a ranking induced, which makes it tractable

$$M(s) := 1 - \frac{1}{Q}\sum_{q=1}^{Q} \frac{1}{Z(y^{(q)})} \sum_{i=1}^{n^{(q)}} (2^{y_i^{(q)}} - 1) A_i^{(q)}(s).$$

The linear model $z = \langle \omega, x \rangle$ of ListNet can be used for s to easily derive the gradient of this quantity and minimize it in a few steps. This procedure is computationally very efficient but relies on a pairwise model $p^{(q)}(\pi \mid s)$.

5.5 Consistency of the NDCG family

We propose to study the consistency of the NDCG family of losses defined by the functions F and G proceeding as Duchi *et al.* (2013, [1]). For a query q, recall the NDCG loss

$$\varphi_{\mathrm{NDCG}}(z, y) = 1 - \frac{1}{Z(y)} \sum_{i=1}^{n} \frac{G(y_i)}{F(j_i)}.$$

We consider a probability distribution μ on the topology of the relevance scores y and define the *pointwise conditional risk* by integrating out the relevance scores. The quantity

$$\int_y \varphi_{\mathrm{NDCG}}(z, y) \mathrm{d}\mu = 1 - \sum_{i=1}^{n} \frac{1}{F(j_i)} \int_y \frac{G(y_i)}{Z(y)} \mathrm{d}\mu$$

is minimized with respect to z when the ranking induced by this score corresponds to the order of the integral vector $\big(\int_y \frac{G(y_i)}{Z(y)} \mathrm{d}\mu\big)$, as shown by Ravikumar *et al.* in [31]. Duchi gives a characterization of the Fisher-consistent surrogates losses for the NDCG family. First we define the vector of scores z corresponding to the order of the integral vector for the distribution

$$Z(\mu) := \left\{ z \in \mathbb{R}^n \;\middle|\; z_i > z_j \text{ when } \int_y \frac{G(y_i)}{Z(y)} \mathrm{d}\mu > \int_y \frac{G(y_j)}{Z(y)} \mathrm{d}\mu \right\}.$$

Duchi's consistency theorem. A surrogate loss φ is Fisher-consistent for the NDCG family if and only if for all query $q \in [\![1, Q]\!]$ and distributions $\mu^{(q)}$,

$$\inf_z \left\{ \varphi(z, \mu^{(q)}) - \inf_{z'} \varphi(z', \mu^{(q)}) \;\middle|\; z \notin Z(\mu^{(q)}) \right\}.$$

An interesting family of losses satisfying this theorem is given by the *anti-kernel* approach. For any function $\phi : \mathbb{R} \to \mathbb{R}_+$ nonincreasing, differentiable and such that $\phi'(0) < 0$, the surrogate defined as

$$\varphi(z, y) = \sum_{i=1}^{n} \frac{G(y_i)}{Z(y)} \sum_{j=1}^{n} \phi(z_j - z_i)$$

is Fisher-consistent by Duchi's theorem. The optimization procedure developed previously corresponds to this family of surrogate losses with the sigmoid function $\phi(\delta) := S_2(-\delta) = \frac{1}{1+e^{2\delta}}$ which satisfies the requirements to ensure consistency.

5.6 Learning with rank aggregation

Following the work of Duchi, we propose to leverage on rank aggregation to produce labels associated with queries. We need to make an assumption on the limiting behaviour of the aggregation scheme $A_L : \mathbb{R}^{n\times L} \to \mathbb{R}^n$ when $L \to \infty$.

Assumption A. For a given query $q \in [\![1, Q]\!]$, let a sequence of measures $m_1^{(q)}, m_2^{(q)}, \cdots$ be drawn i.i.d. conditional on q. For L measures, we define the aggregated score $Y_L := A_L(m_1^{(q)}, \cdots, m_L^{(q)})$ and denote $\mu_L^{(q)}$ its distribution. There exists a limiting distribution $\mu^{(q)}$ such that

$$\mu_L^{(q)} \xrightarrow[L\to\infty]{\text{law}} \mu^{(q)}.$$

We define a class of estimators taking advantage of the weak convergence of Assumption A. Duchi proposed to use U-statistics to develop a class of statistical procedures ensuring convergence to the empirical risk R^{φ}. We build a surrogate empirical risk using the robustness of the U-statistic in its loss, which depends on n and the order of the U-statistic $u < L$

$$\hat{R}^{\varphi}_{L,u}(s) := \frac{1}{Q}\sum_{q=1}^{Q}\binom{L}{u}^{-1}\sum_{l_1<\cdots<l_u}\varphi(s(x^{(q)}), A(m_{l_1}^{(q)}, \cdots, m_{l_u}^{(q)})).$$

This is an unbiased estimator of the surrogate population risk

$$R^{\varphi}_{L,u}(s) := \frac{1}{Q}\sum_{q=1}^{Q}\mathbb{E}^{(q)}\left[\varphi(s(x^{(q)}), A(m_1^{(q)}, \cdots, m_u^{(q)}))\right]$$

where $\mathbb{E}^{(q)}$ denotes the expectation under $\mu^{(q)}$. Furthermore, note that when $u \to \infty$, $R^{\varphi}_{L,u}(s)$ tends to the *complete aggregation* surrogate empirical risk

$$\frac{1}{Q}\sum_{q=1}^{Q}\varphi(s(x^{(q)}), A(\{m_l^{(q)} \mid l \in [\![1, L]\!]\})).$$

We now enumerate sufficient conditions for the estimator to converge in probability. By sorting the query indexes, we assume without loss of generality that $\mathbb{P}(q)$ the probability of observing query q is nonincreasing in the associated index of q. The following assumption describes the tail of the query distribution.

Assumption B. There exist a constant $\beta > 0$ such that $\forall q \in [\![1, Q]\!], \mathbb{P}(q) = \mathcal{O}(q^{-\beta-1})$ where q denotes the query index.

This assumption allows to deal with infinite sets of queries while being automatically satisfied for finite sets. The following assumption concerns the regularity of the loss function φ.

Assumption C. The loss φ is bounded and Lipschitz continuous over the set of ranking functions $h : \mathcal{A} \to \mathfrak{S}_n$ generated by the L measures. For all query $q \in [\![1, Q]\!]$, there exist constants B_L and K_L such that

$$\forall y \in \mathbb{R}^{n^{(q)}}, \begin{cases} \forall s : \mathbb{R}^L \to \mathbb{R},\ 0 \leq \varphi(s(x^{(q)}), y) \leq B_L \\ \forall s_1, s_2 : \mathbb{R}^L \to \mathbb{R},\ \left|\varphi(s_1(x^{(q)}), y) - \varphi(s_2(x^{(q)}), y)\right| \leq K_L\|s_1 - s_2\| \end{cases}.$$

It is sufficient to choose $\varphi(\cdot, y)$ convex to satisfy assumption C. In order to guarantee uniform convergence, we let the aggregation order u_L grows with the number of measures L. The next assumption imposes regularity constraints on both the loss function φ and the aggregation scheme A.

Assumption D. There exist constants C, $\rho > 0$ such that for all $q \in [\![1, Q]\!]$, $u_L \in \mathbb{N}$, $B_L = o(u_L^{\rho})$, $s : \mathbb{R}^L \to \mathbb{R}$,

$$\left|\mathbb{E}^{(q)}\left[\varphi(s(x^{(q)}), A(m_1^{(q)}, \cdots, m_u^{(q)}))\right] - \lim_{u' \to \infty} \mathbb{E}^{(q)}\left[\varphi(s(x^{(q)}), A(m_1^{(q)}, \cdots, m_{u'}^{(q)}))\right]\right| \leq C B_L u_L^{-\rho}.$$

Duchi's convergence theorem. Provided a few topological constraints on the structure of the ranking functions, assumptions B, C and D insure the convergence

$$\sup_{s : \mathbb{R}^L \to \mathbb{R}} \left|\hat{R}_{L,u}^{\varphi}(s) - R_{L,u}^{\varphi}(s)\right| \xrightarrow[L \to \infty]{} 0.$$

Therefore, by using a Fisher-consistent and convex loss, this theorem gives a statistical procedure that is both computationally tractable and asymptotically consistent. Duchi's convergence result is the theoretical basis of learning to rank using aggregation as supervision. The framework developed allows to tackle a wide variety of real world problems.

Algorithm 7 Stochastic U-statistics NDCG optimization

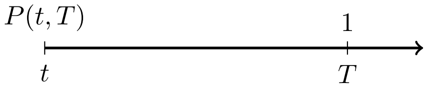
```
procedure NdcgUstat($u, \lambda, \eta$)
    $\omega \leftarrow \mathcal{N}(0, \epsilon)$                                      ▷ randomly initialize the model
    for $t \in [\![1, T]\!]$ do                                        ▷ $T$ iterations
        $q \sim \mathcal{U}_{[\![1,Q]\!]}$                                           ▷ sample a query
        for $i \in [\![1, n^{(q)}]\!]$ do
            compute the prediction $z^{(q)} = \langle \omega, x^{(q)} \rangle$
            compute the gradient $\nabla\varphi_{\text{NDCG}} := \nabla_\omega \varphi_{\text{NDCG}}(A(m_{l_1}^{(q)}, \cdots, m_{l_u}^{(q)}), z^{(q)})$
            compute the regularized gradient $\nabla\varphi := \nabla\varphi_{\text{NDCG}} + \lambda\omega$    ▷ Tychonoff regularization
            update the model $\omega \leftarrow \omega - \eta\nabla\varphi$
    return $\omega$                                                    ▷ return learned model
```

6 Bond portfolio selection for the yield curve fitting problem

Understanding the term structure of interest rates is central in finance and economics since it allows both pricing financial instruments and understanding economic conditions. It is formalized by the estimation of the yield curve from a subset of available money market instruments, the *yield curve fitting* problem. The issue is critical for the good performance of money markets and is therefore addressed by all central banks, offering a multitude of fitting methods. Nonetheless, the selection of the instruments is often carried out manually and stability issues make it challenging. We propose to apply aggregation-based ranking systems to automate the selection of bonds to build the yield curve.

6.1 Coupon bonds and the term structure of the interest rates

Following the notations of Filipović (2005, [33]), we consider today's time t and a maturity date $T > t$. The term structure can be described by the value $P(t, T)$ at time t of one unit of currency at time T. The virtual contract associated is called *zero-coupon bond* (ZCB) with maturity date T. We are looking to estimate the function $T \mapsto P(t, T)$ for a given time t.



Market hypothesis. There exists a frictionless market for ZCB for all maturities T and this market is efficient, *ie.* without arbitrage opportunities.

These assumptions are required to verify the *expectations hypothesis* which states that the long-term rate is purely determined by current and future expected short-term rates. Hence, the expected final wealth from investing in a sequence of short-term instruments equals the final wealth from investing in long-term instruments.

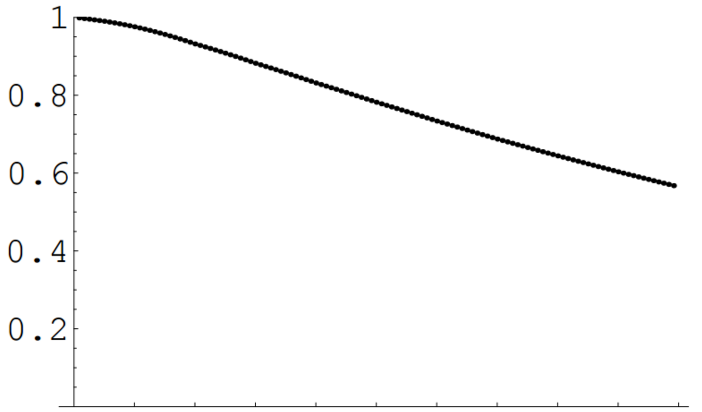


Figure 5: Example of curve $T \mapsto P(t, T)$ of the US Treasury Bonds market *T-bills* for t = march 2002, time axis graduated in years.

The following property can be proved by a no-arbitrage argument and imposes regularity constraints on the curve, $\forall t \leq T \leq S,\ P(t, S) = P(t, T)P(T, S)$, hence the smooth appearance of the function $T \mapsto P(t, T)$.

We first define the *continuously compounded spot rate* $R(t,T) := -\frac{\log P(t,T)}{T-t}$ to build the usual *instantaneous short rate* at time t by taking the limit $r(t) := \lim_{T\to t^+} R(t,T)$. Therefore, for a small interval Δt, we have the first order approximation of the discount factor

$$\frac{1}{P(0,\Delta t)} = 1 + r(0)\Delta t + o(\Delta t).$$

For a coupon-paying bond with cashflow $\{c_i\}_{i=1}^{m-1}$ at times $\{T_i\}_{i=1}^{m-1}$, maturity T_m and notional N, the expected price at time t writes as

$$p(t) = \sum_{i=1}^{m} P(t,T_i)c_i + P(t,T_m)N.$$

The term structure of interest rates is characterized by the knowledge of the ZCB $P(t,T)$ or equivalently by the yield curve $T \mapsto R(t,T)$ such that $P(t,T) = e^{-R(t,T)(T-t)}$. The yield corresponds to the rate explaining the observed market prices $p(t)$. Similarly, we can compute the yield of a coupon-paying bond with the appropriate frequency for discounting the coupons. Note that we only use riskless, government bonds to ensure that $P(t,\cdot)$ remains under 1, *ie* respecting the *time value of money* principle stating that one unit of currency is worth more now than the same amount in the future.

6.2 Nelson–Siegel–Svensson model

The Nelson-Siegel model has been introduced in 1987 to propose a global model to the term structure of interest rates. It is defined on the yields and can be fitted using market yields. With the extension due to Svensson (1994), the Nelson–Siegel–Svensson (NSS) model became widely used by central banks for its ability to explain the term structure. We call $y(t)$ the market yield at time t, the unicity is insured by the market hypothesis.

The (NSS) model is defined on the yields

$$\hat{y}_{\vec{\beta},\vec{\lambda}}(t) = \underbrace{\beta_1}_{\substack{\text{long-run}\\ \text{yield level}}} + \underbrace{\beta_2 \frac{1-\exp\frac{-t}{\lambda_1}}{\frac{t}{\lambda_1}}}_{\text{exponential decay on short-end}} \underbrace{+\beta_3 \left(\frac{1-\exp\frac{-t}{\lambda_1}}{\frac{t}{\lambda_1}} - \exp\frac{-t}{\lambda_1}\right)}_{\text{first hump}} \underbrace{+\beta_4 \left(\frac{1-\exp\frac{-t}{\lambda_2}}{\frac{t}{\lambda_2}} - \exp\frac{-t}{\lambda_2}\right)}_{\text{second hump}}.$$

Calibration of the model requires to learn the yields parameters $\vec{\beta} = (\beta_1,\beta_2,\beta_3,\beta_4)$ and the time-homogeneous parameters $\vec{\lambda} = (\lambda_1,\lambda_2)$. This model allows to fit different behaviour of the yield curve on both ends and two *humps*. Svensson improvement consisted into addingthe second hump and the corresponding parameters β_4, λ_2 to widen the fitting power of the model. This structure has been proposed following a principal component analysis of historical market data.

The fitting is carried out my minimizing a mean squares metric between the predicted yields and market-observed yields $\{y(t_i)\}_{i=1}^{m}$, which allows the use of a wide variety of different instruments. The (NSS) *fitting problem* writes as

$$(\vec{\beta}^*, \vec{\lambda}^*) = \arg\min_{\vec{\beta},\vec{\lambda}} \sum_{i=1}^{m} \left(\hat{y}_{\vec{\beta},\vec{\lambda}}(t_i) - y(t_i)\right)^2. \tag{NSS fitting}$$

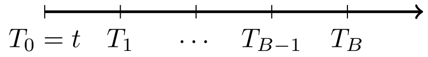
The most usual calibration method relies on the Nelder-Mead algorithm, a gradient-free optimization method useful for multidimensional problems. Gilli *et al.* (2010, [34]) demonstrated the stability issues associated with the parametric problem (NSS fitting): the objective function is not convex and exhibits several local minima. If the fitting's stability can be improved by carrying several runs, the selection of the bond portfolio remains challenging and is often achieved by trial and error. Automating the process of bond selection based on characteristic features would offer a clear advantage the manual selection.

6.3 Maturity buckets and features

For a given date t, we have a set of available riskfree bonds $\mathcal{B}_t$. The bond portfolio selection problem consists into choosing a portfolio $S_t \subseteq \mathcal{B}_t$ to fit the yield curve with the (NSS) model. For our analysis, each bond is characterized by features capturing different properties of the financial instrument. For bond $i \in \mathcal{B}_t$, we propose the following characteristics.

- Coupon C_i: amount of coupon paid. The frequency of the cashflow can be incorporated by bringing all the coupons to the same virtual frequency.
- Issue volume IV_i: volume of bond traded during the day. Used as a liquidity measure for the instrument.
- Modified duration D_i: value associated to the cashflow describing the price sensitivity of the bond. It can be seen as a derivative with respect to yield.
- Yield y_i: amount of return realized by the investor on the bond. Immediately after issuance, it corresponds to the coupon rate but then varies on the opposite direction than the price.

Furthermore, the selection imposes maturity constraints to avoid selection too many bonds on the same side of the curve. We deal with this issue by dividing the time scale into *maturity buckets* and selecting bonds in each of the buckets. This choice allows to provide the fitting method with enough yield information on each part of the curve and avoid the unstable extrapolation on the long end. The selected buckets are defined between the *pillars* $\mathbb{T} = \{T_1, \cdots, T_B\} = \{$today, 1, 3, 6 months, 1, 3, 5, 10, 15, 20, 30 years$\}$. For a pillar T_b, the corresponding bucket is $[T_b, T_{b+1}[$ or $[T_B, +\infty[$ for the last one.



The bond space and the selection S_t are decomposed on the maturity buckets as $\mathcal{B}_t = \bigsqcup_{b \in [\![1,B]\!]} \mathcal{B}_{t,b}$ and $S_t = \bigsqcup_{b \in [\![1,B]\!]} S_{t,b}$. A vector of features is associated to the bond portfolio S_t by aggregating each characteristic feature of the bonds per bucket. For example, the *coupon arithmetic mean* of bucket b is defined as

$$\forall t \in [0, +\infty[, b \in [\![1, B]\!],\ AC_{t,b} := \frac{1}{|S_{t,b}|} \sum_{i \in S_{t,b}} C_i.$$

And the *coupon maximum gap* of bucket b is defined as

$$\forall t \in [0, +\infty[, b \in [\![1, B]\!],\ GC_{t,b} := \max_{i \in S_{t,b}} C_i - \min_{i \in S_{t,b}} C_i.$$

6.4 Bond portfolio selection

The selection procedure hence selects a given number of bonds in each bucket, which breaks the combinatorial complexity of the selection problem. The pillars can be considered metaparameters of the learning model. The task of selection requires to remove outlier bonds with extreme features which might induce fitting errors. We can apply filters on these aggregated values to remove outliers in a dynamic fashion. It provides the user with a query interface to manually tune the sensitivity of the selection to each feature.

The most robust filters are conditions on quantiles. For a given date t, pillar b and threshold $0 < \alpha < 1$, we denote the empirical α-quantile on all possible subsets $S_{t,b}$ by

$$q_{AC_{t,b}}(\alpha) := \widehat{F_{AC_{t,b}}}^{-1}(\alpha).$$

It allows to write a query as a sequence of filters of the form $\big(AC_{t,b} \leq q_{AC_{t,b}}(95\%)\big)$ or $\big(GC_{t,b} \leq q_{GC_{t,b}}(75\%)\big)$ and so on to restrict the space of alternatives. To fully exploit the potential of query-based learning to rank algorithms, the features of bond portfolios should depend on the preferences of the user, *eg.* increasing the precision on one end of the curve.

The fitting can be judged by out-of-sampling metrics in each maturity bucket, *ie.* comparing the curve with bonds not used for the parameters calibration. This property of the problem makes it adapted to the learning to rank framework since the *iid hypothesis* of these out-of-sampling metrics can be decently assumed if the bucket granularity is finer than the fitting model. For example, in the case of the (NSS) model, we need to have at least a bucket for each end of the curve and for each hump. A finer bucket selection would reduce the number of possible bond portfolios but would require more manual tuning. Therefore, we face a trade-off between too much tuning on the choice of buckets and the minimum imposed by the term structure model.

We consider the *root-mean-square error* between the predicted curve and the all the available bonds on each bucket $\mathcal{B}_b$,

$$\forall b \in [\![1, B]\!],\ RMSE_b(\vec{\beta}, \vec{\lambda}) := \sqrt{MSE_b(\vec{\beta}, \vec{\lambda})} = \sqrt{\mathbb{E}_{t\sim\mathcal{B}_b}\Big(\hat{y}_{\vec{\beta},\vec{\lambda}}(t) - y(t)\Big)^2}$$

and we estimate the RMSE with the out-of-sample bonds to obtain the metric

$$\forall b \in [\![1, B]\!],\ \widehat{RMSE}_b(\vec{\beta}, \vec{\lambda}) := \sqrt{\sum_{i\in\mathcal{B}_{t,b}} \big(\hat{y}_{\vec{\beta},\vec{\lambda}}(t_i) - y(t_i)\big)^2}.$$

For a given date t, query q and bond portfolio S_t corresponding to q, we can build a vector of features in $\mathbb{R}^d$ with characteristics like $AC_{t,b}$, $GC_{t,b}$ when $b \in [\![1, B]\!]$. The algorithm aggregates the RMSE errors on each bucket so that $L = B$ and $m_b^{(q)} := \widehat{RMSE}_b(\vec{\beta}, \vec{\lambda})$. We use the Tychonoff method $\frac{\lambda}{2}\|\omega\|_2^2$ to regularize the objective surrogate. Simple heuristics on the choice of the *learning rate* η can be used, like dividing by 10 the learning rate in case of successful step, *ie.* $\eta \leftarrow \frac{\eta}{10}$. We test the convergence on the l_2-residual, which gives us a stopping criterion

$$l_2 := \sqrt{\sum_{i=1}^{d} \big(\eta\nabla\varphi_i\big)^2}.$$

A model $\omega^* \in \mathbb{R}^d$ can be learned from historical data with the stochastic U-statistics NDCG optimization routine. Given a query q, we can then predict the ranking of a given set of bond portfolios $\{S_t^{(i)}\}_{i=1}^n$ by computing their vectors of features $\{x_i\}_{i=1}^n$ and sorting the predicted scores $\langle\omega^*, x_i\rangle$. Therefore, the algorithm automates the selection of the bond portfolio used for the fitting, a task which is usually carried by hand. Furthermore, the user can provide a set of filters as query to impose constraints on the selection of bonds and therefore on the fitting itself.

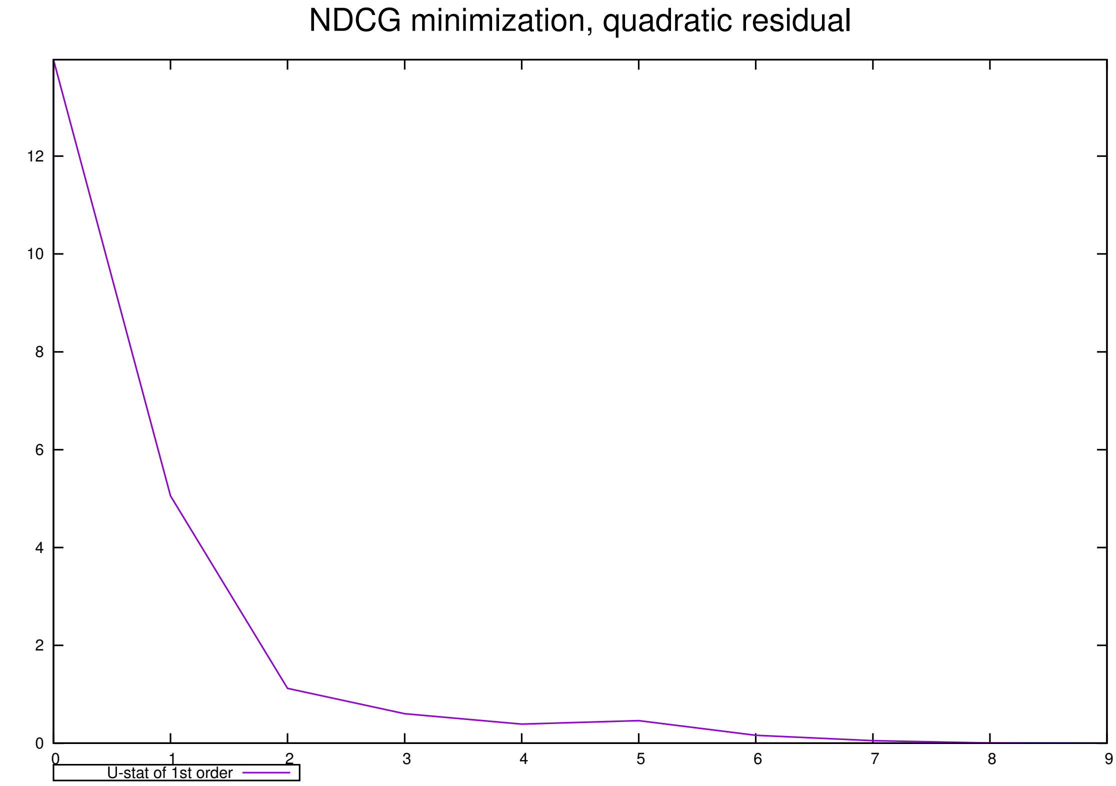


Figure 6: Residual l_2 of the gradient descent as a function of the number of iterations. The residual has been stabilized by averaging on several runs of the stochastic gradient descent. Implementation in C++11/Boost, runtime of approximately 15 seconds on Linux Ubuntu Intel® Xeon® 16 CPUs X5550 @2.67GHz. Dataset of german government bonds, $Q = 251$, $B = 6$, $d = 48$, $L = 36$, $n = 3998$ query/portfolio pairs. Years 2015/2016 of data from the *Bundesrepublik Deutschland Finanzagentur GmbH.*

7 Conclusion

We carried out an axiomatic analysis of ranking systems by enumerating the desirable properties and linking several results of recent papers. Furthermore, using the convergence results listed by Duchi, we proposed a learning framework adapted to problems where independent and identically distributed metrics can be provided. We gave an extensive description of a wide variety of aggregation schemes to aggregate these metrics and provide supervision to train models. In the case of a linear model $\langle \omega, \cdot \rangle$, we disclosed two learning to rank algorithms, respectively based on the Cross Entropy surrogate and the NDCG family of surrogates. Finally we offered an automated selection procedure to tackle the problem of bond portfolio selection to fit the yield curve, using the convergence results of Duchi on financial time series data.

References

[1] J.C. Duchi, L. Mackey, M.I. Jordan *The Asymptotics of Ranking Algorithms* The Annals of Statistics 41:5, 2013

[2] A. Prasad, H. Pareek, P. Ravikumar *Distributional Rank Aggregation, and an Axiomatic Analysis.* Department of Computer Science, The University of Texas, Austin, 2015

[3] M.J. marquis de Condorcet *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix.* Imprimerie royale, 1785

[4] K.J. Arrow *Social choice and individual values.* Wiley, New York, 1951

[5] H.A. Soufiani, D.C. Parkes, L. Xia. *A Statistical Decision-Theoretic Framework for Social Choice.* In Advances in Neural Information Processing Systems, pp. 3185–3193, 2014.

[6] R.D. Luce *Luce's choice axiom* Scholarpedia 3(12):8077, 2008

[7] J.J. Bartholdi, C.A. Tovey, M.A. Trick *The Computational Difficulty of Manipulating an Election.* School of Industrial and System Engineering, Georgia Institute of Technology, 1989

[8] R.Y. Rubinstein, D.P. Kroese *The Cross-Entropy Method. A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning.* Information Science & Statistics, 4(2):132, 2004

[9] D.P. Kroese, R.Y. Rubinstein, P.W. Glynn *The Cross-Entropy Method for Estimation.* The University of Queensland, Brisbane, 2013

[10] S. Lin, J. Ding *Integration of Ranked Lists via Cross Entropy Monte Carlo with Applications to mRNA and microRNA Studies.* The Ohio State University, 2010

[11] A.J. Walker *New fast method for generating discrete random numbers with arbitrary frequency distributions.* Electronics Letters 10(8):127, 1974

[12] L. Devroye *Non-Uniform Random Variate Generation.* Springer-Verlag, New York, 3:107, 1986

[13] S. Clémençon, J. Jakubowicz, E. Sibony *Multiresolution analysis of incomplete rankings.* ArXiv e-prints, 2014

[14] M. Truchon *An Extension of the Condorcet Criterion and Kemeny Orders.* CRÉFA and Département d'économique. Université Laval, 1998

[15] M. Drissi, M. Truchon *Maximum Likelihood Approach to Vote Aggregation with Variable Probabilities.* CIRPÉE and Département d'économique, Université Laval, 2002

[16] H.P. Young, A. Levenglick *A Consistent Extension of Condorcet's Election Principle.* IIASA Research Report, 1977

[17] C.L. Mallows *Non-null ranking models.* Biometrika 44:114, 1957

[18] J-P. Doignon, A. Pekec, M. Regenwetter *The repeated insertion model for rankings: Missing link between two subset choice models.* Psychometrika 69(1):33, 2004

[19] T. Lu, C. Boutilier *Effective Sampling and Learning for Mallows Models with Pairwise-Preference Data* Journal of Machine Learning Research 15, 2014

[20] R.A. Bradley, M.E. Terry *The rank analysis of incomplete block designs. I. The method of paired comparisons.* Biometrika, 1952

[21] F. Mosteller, L.L. Thurstone *Remark on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations.* Psychometrika, 1951

[22] T.L. Saaty *Decision-making with the AHP: Why is the principal eigenvector necessary.* European J. Oper, 2003

[23] L. Maystre, M. Grossglauser *Fast and Accurate Inference of Plackett–Luce Models.* Proceedings of the Annual Conference on Neural Information Processing Systems, Montreal, 2015

[24] L.R. Ford *Solution of a ranking problem from binary comparisons.* The American Mathematical Monthly, 64(8):28–33, 1957

[25] H.A. Soufiani, D.C. Parkes, L. Xia *Random Utility Theory for Social Choice.* Proceedings of the Annual Conference on Neural Information Processing Systems, Lake Tahoe, 2012

[26] T. Liu *LETOR - Learning to Rank for Information Retrieval* Microsoft Research Asia, 2008

[27] F. Xia, T. Liu, J. Wang, W. Zhang, H. Li *Listwise Approach to Learning to Rank - Theory and Algorithm* Proceedings of the 25th International Conference on Machine Learning, Helsinki, 2008

[28] F. Pedregosa, F. Bach, A. Gramfort *On the Consistency of Ordinal Regression Methods* Institut Europlace de Finance, Louis Bachelier, 2014

[29] Z. Cao, T. Qin, T. Liu, M. Tsai, H. Li *Learning to Rank: From Pairwise Approach to Listwise Approach* Proceedings of the 24th International Conference on Machine Learning, Corvallis, 2007

[30] T. Luo, D. Wang, R. Liu, Y. Pan *Stochastic Top-k ListNet* Tsinghua National Lab for Information Science and Technology, Beijing, 2015

[31] P. Ravikumar, A. Tewari, E. Yang *On NDCG Consistency of Listwise Ranking Methods* University of Michigan, 2011

[32] H. Valizadegan, R. Jin, R. Zhang, J. Mao *Learning to Rank by Optimizing NDCG Measure* Advances in Neural Information Processing Systems 22, 2009

[33] D. Filipović *Interest Rate Models.* University of Munich, 2005

[34] M. Gilli, S. Große, E. Schumann *Calibrating the Nelson–Siegel–Svensson model.* COMISEF working papers series, 2010.