A Non-Utilitarian Discrete Choice Model for Preference Aggregation

Martin Durand[0000-0001-9117-8661], Fanny Pascual[0000-0003-0215-409X], and Olivier Spanjaard[0000-0002-9948-090X]

Sorbonne Université, LIP6, CNRS, France
{martin.durand,fanny.pascual,olivier.spanjaard}@lip6.fr

Abstract. We study in this paper a non-utilitarian discrete choice model for preference aggregation. Unlike the Plackett-Luce model, this model is not based on the assignment of utility values to alternatives, but on probabilities p_i to choose the best alternative (according to a ground truth ranking r^*) in a subset of i alternatives. We consider $k-1$ parameters p_i (for $i=2$ to k) in the model, where k is bounded by the number m of alternatives. We study the application of this model to voting, where we assume that the input is a set of choice functions provided by voters. If $k=2$, our model amounts to the model used by Young [25] in his statistical analysis of Condorcet's voting method, and a maximum likelihood ranking is a consensus ranking for the Kemeny rule (1959). If $k>2$, we show that, under some restrictive assumptions about probabilities p_i, the maximum likelihood ranking is a consensus ranking for the k-wise Kemeny rule [10]. In the general case, we provide a characterization result for the maximum likelihood ranking r and probabilities p_i. We propose an exact and a heuristic algorithm to compute both ranking r and probabilities p_i. Numerical tests are presented to assess the efficiency of these algorithms, and measure the model fitness on synthetic and real data.

1 Introduction

Preference aggregation is ubiquitous in multiple fields, among which are social choice [2, 22], information retrieval [7], collaborative filtering [17], or peer grading [20]. The aggregation problem is formulated as follows: given n agents (or voters) and m alternatives (or candidates), each agent's preferences are specified by a ranking (permutation) of the alternatives, and the aim is to determine a single *consensus* ranking. Alternatively, preferences can also be expressed as choice functions instead of rankings [1], i.e., each agent chooses her preferred candidate among various subsets of candidates. A choice function allows more possibilities for the voters (cyclic preferences are even possible), and may be easier to elicit if only a few subsets of candidates are considered. However, if all subsets of candidates are considered, their number becomes quickly very large (2^m). The procedure producing a consensus ranking from the n agents' preferences (expressed as rankings or choice functions) is called a *voting rule*.

A stream of research aims to rationalize voting rules by using statistical models for rank data, whose characteristics depend on the application domain (see e.g. [24]). This assumption of a statistical model behind the agents' preferences dates back to Condorcet. As emphasized by Young [25], "Condorcet argued that if the object of voting is to determine the 'best' decision for society but voters sometimes make mistakes in their judgments, then the majority alternative (if it exists) is statistically most likely to be the best choice."

Young's examination of Condorcet's work through the lens of modern statistics leads him to put forward the Kemeny rule [12]. This well-known rule consists of producing a consensus ranking r that minimizes the number of disagreements between r and the pairwise preferences of the agents on the candidates. Young shows that a consensus ranking for the Kemeny rule is a Maximum Likelihood Estimate (MLE) of a "ground truth" ranking r^* of the alternatives if one assumes that the pairwise preferences of the voters follow a statistical model parameterized by r^* under specific assumptions. The assumptions (already made by Condorcet) are: 1) in every pairwise comparison, each voter chooses the better alternative in r^* with some fixed probability p, with $p > \frac{1}{2}$; 2) each voter's judgment on every pair of alternatives is independent of her judgment on every other pair[1]; 3) each voter's judgment is independent of the other voters' judgments.

When voters' preferences are expressed as rankings, it is also known that a consensus ranking for the Kemeny rule is an MLE of a ground truth ranking r^* for a distance-based statistical model for ranking data [5]. Consider indeed the conditional probability distribution Pr on rankings r' of candidates defined by $Pr(r'|r^*) \propto 2^{-\delta(r^*,r')}$, where $\delta(r^*, r')$ is the Kendall tau distance between r^* and r' (number of pairwise disagreements between r^* and r'). Assuming that each voter's judgment is independent of the other voters' judgments, it is easy to show that the Kemeny rule returns a ranking r maximizing $Pr(r_1, \ldots, r_n|r) = \prod_{j=1}^{n} Pr(r_j|r) = 2^{-\sum_j \delta(r,r_j)}$, i.e., an MLE of r^*.

Other works about the use of MLE for preference aggregation explore the estimation of the parameters of *discrete choice models* from voting data. A discrete choice model consists of predicting the probabilities, called *choice probabilities*, of choosing $c \in S$ when presented with a subset S of alternatives, for each possible subset S [14]. A set of agents' rankings can be seen as choice data by considering that each ranking rationalizes a *choice function*. A choice function f picks a favorite alternative in any subset S of alternatives. For instance, the ranking $1 \succ 2 \succ 3$ (where "$\succ$" stands for "is preferred to") rationalizes the choice function $f(\{1,2\}) = 1$, $f(\{1,3\}) = 1$, $f(\{2,3\}) = 2$, and $f(\{1,2,3\}) = 1$. The most famous discrete choice model is due to Plackett-Luce. It consists of assigning a utility u_c to each alternative c, and setting the probability $Pr(f(S) = c)$ to choose c in S equal to $u_c / \sum_{d \in S} u_d$. The corresponding voting rule returns the ranking of alternatives by decreasing order of maximum likelihood utilities. Unlike most discrete choice model, the model we propose hereafter does not rely on the assignment of utility values (or utility distributions) to alternatives. Like the Plackett-Luce model, and unlike the model we will propose and study, most discrete choice models rely on the assignment of utility values (or utility distributions) to alternatives.

[1] Note that this assumption allows the preferences to be cyclic.

The use of discrete choice models based on utilities for preference aggregation deviates from Young's point of view. Indeed, Young uses distinct parameters to model, *on the one hand*, the respective "objective" skills of the candidates, namely the parameter r^* (ground truth ranking), and *on the other hand*, the "reliability" of the judgments of the voters, namely the parameter p (the closer the probability p is to 1, the more consistent the preferences are with the ground truth ranking). In discrete choice models based on utilities, the utilities are used *both* for modeling the objective skills of the candidates and the reliability of the judgments (the greater the differences in utilities, the more reliable the voters' judgments). Besides, unlike Young's model, that is related to the Kemeny rule, the consensus rankings obtained by sorting the candidates by decreasing order of maximum likelihood utilities are not related to well-identified voting rules.

Our contributions. We propose a discrete choice model inspired by Young's model for the Kemeny rule. Given a ground truth ranking r^* of the alternatives, the choice of an agent in a subset of i alternatives is consistent with r^* with a probability p_i (p_i is $\alpha_i > 1$ times greater than the probability to choose any other given candidate in a subset of size i). Unlike many discrete choice models used for social choice, the model is thus *non-utilitarian*, i.e., not based on the assignment of utility scores to alternatives. While the introduction of utility scores is appealing because the cardinal data are richer than the ordinal ones, the interpretation of such utility scores is not always obvious, e.g., when comparing artworks. We show the following results regarding the model we propose:

– Proposition 1 states that, if the value of α_i does not depend on i, then a maximum likelihood ranking is a consensus ranking for the k-wise Kemeny rule, a recently introduced voting rule [10].
– If values α_i depend on i, we provide a characterization result (Proposition 2) for a maximum likelihood estimation of the ground truth ranking r^* and α_i's. The characterization involves a weighted variant of the k-wise Kemeny rule.
– Based on Proposition 2, we provide an exact algorithm and a heuristic algorithm for determining a maximum likelihood couple in the general case.
– Finally, using synthetic and real data, we present numerical tests to assess the efficiency of these algorithms, as well as the model fitness to data.

2 Related work

The related work concerns either the maximum likelihood approach to voting, or set extensions of the Kemeny rule.

The maximum likelihood approach to voting. In this approach, we make the assumption that a true "objective" ranking of the candidates exists, and that the preferences expressed by the voters are noisy observations of this true ranking. If the preferences are rankings drawn i.i.d. from a distribution, the probability of observing a set $\mathcal{P} = \{r_1, \ldots, r_n\}$ is then $Pr(\mathcal{P}|r) = \prod_{j=1}^{n} Pr(r_j|r)$. Each probability model for $Pr(r_j|r)$ induces a voting rule where a ranking maximizing

$Pr(\mathcal{P}|r)$ (the likelihood) is a consensus ranking. Drissi-Bakhkhat and Truchon [8] investigate a setting in which the probability of comparing two alternatives consistently with a ground truth ranking r^* is increasing with the distance between them in r^*. This leads to a new voting rule that the authors examined from an axiomatic point of view. While every noise model on the votes[2] induces a voting rule, Conitzer and Sandholm [6] study the opposite direction, using it as a way to rationalize voting rules. They identify noise models for which an MLE ranking is a consensus ranking of well-known voting rules (scoring rules and single transferable vote), and on the contrary, for other rules (Bucklin, Copeland, maximin), they show that no such noise model can be constructed. Conitzer et al. [5] pursue this line of work, providing an exact characterization of the class of voting rules for which a noise model can be constructed. More recently, Caragiannis et al. [4] study how many votes are needed by a voting rule to reconstruct the true ranking. Another line of research focuses on the use of discrete choice models in social choice. Soufiani et al. [23] study an extension of the Plackett-Luce model. This model can be viewed as a random utility model in which the utilities of alternatives are drawn i.i.d. from a Gumbel distribution. They propose a random utility model based on distributions in the exponential family (to which Gumbel distributions belong), as well as inference methods for the parameters.

Set extensions of the Kemeny rule. Gilbert et al. [10] introduce the k-wise Kemeny rule, show that the computation of a consensus ranking according to this rule is NP-hard, and provide a dynamic programming procedure for this purpose. At least two other set extensions of the Kemeny rule have been proposed. Both extensions consider a setting in which, although the voters have preferences over a set $\mathcal{C}$, the election will in fact occur on a subset $S \subseteq \mathcal{C}$ drawn according to a probability distribution on $2^{\mathcal{C}}$ [3, 13]. A consensus ranking r is then one that minimizes, in expectation, the number of voters' disagreements with the chosen candidate in S (a voter disagrees with r on S if $t_r(S)$ is not her most preferred candidate in S). Baldiga and Green study a setting in which the probability $Pr(S)$ only depends on the cardinality of S. Lu and Boutilier study a special case of the previous setting, where each candidate is unavailable in S with a probability p, independently of the others, i.e., $Pr(S) = p^{|C \setminus S|}(1-p)^{|S|}$. Proposition 2 later in the paper uses a weighted sum of disagreements δ_α^k on subsets of size at most k that is formally equivalent to the rule used by Baldiga and Green for $k = m$: the weights $\log \alpha_i$ assigned to disagreements on subsets S of size $i = |S|$ play the role of $Pr(S)$. However, the viewpoint we take here is completely different, as the values α_i are not given, but inferred from the choice data. In addition, to determine a maximum likelihood ranking for our model, we do not minimize δ_α^k only, but the sum of δ_α^k and another term.

[2] When the votes are viewed as noisy perceptions of a ground truth ranking r^*, a noise model is the mathematical description of the probabilities of the votes based on r^*.

3 Preliminaries

In the following, we will consider that the preferences of the agents are expressed as choice functions. A first possibility to elicit these choice functions is by asking each agent to give her most preferred alternative for each subset of size at most k – this may be a good solution if there are few candidates and k is not too large, or when the agents are not able to give their preferences as rankings. Another possibility is to ask for rankings, and infer choice functions from them (the choice in a subset S of candidates is the highest ranked candidate among S) – a ranking can be seen as a compact representation of a choice function.

Example 1. Let us consider 3 candidates $\{c_1, c_2, c_3\}$ and 10 voters with preferences, expressed as rankings, as follows: 3 voters of type I have preferences $c_1 \succ c_2 \succ c_3$; 3 voters of type II have preferences $c_3 \succ c_1 \succ c_2$; 2 voters of type III have preferences $c_2 \succ c_1 \succ c_3$; 2 voters of type IV have preferences $c_3 \succ c_2 \succ c_1$.
This preference profile yields the choice function profile given in Table 1, where each cell gives the favorite alternative $f_j(S)$ in S for voter j of the type corresponding to the row. Considering the rightmost column, one sees that c_1 (resp. c_3) is the preferred candidate in $\{c_1, c_2, c_3\}$ for voters of type I (resp. II and IV).

	$S=\{c_1,c_2\}$	$S=\{c_1,c_3\}$	$S=\{c_2,c_3\}$	$S=\{c_1,c_2,c_3\}$
$f_j(S)$ (j of type I)	c_1	c_1	c_2	c_1
$f_j(S)$ (j of type II)	c_1	c_3	c_3	c_3
$f_j(S)$ (j of type III)	c_2	c_1	c_2	c_2
$f_j(S)$ (j of type IV)	c_2	c_3	c_3	c_3

Table 1. The choice function profile in Example 1.

Let $\mathcal{V}=\{1,\ldots,n\}$ be a set of n agents (or voters) and $\mathcal{C}$ a set of m alternatives (or candidates). We denote by R the set of the $m!$ possible rankings of $\mathcal{C}$. For $k \in \{2,\ldots,m\}$, we denote by Δ_k the set of all subsets S of $\mathcal{C}$ such that $2 \leq |S| \leq k$. Given a value $k \in \{2,\ldots,m\}$, each agent $j \in \mathcal{V}$ has a choice function $f_j : \Delta_k \to \mathcal{C}$ which gives, for each subset S of alternatives of size at most k, her preferred alternative in S (assuming that each agent has only one favorite alternative per subset). We denote by $\mathcal{F}_k$ the set of all possible choice functions on sets of size at most k. A *choice function profile* $\mathcal{P}=(f_1,\ldots,f_n) \in \mathcal{F}_k^n$ is a tuple of n choice functions f_j, one per agent. In this setting, the purpose of preference aggregation is to determine a *consensus ranking* from the choice functions in $\mathcal{P}$. A voting rule $\mathcal{R}: \mathcal{F}_k^n \to (2^R \setminus \{\emptyset\})$ in which ballots are choice functions, maps each choice function profile to a non-empty set of consensus rankings.

The statistical model for choice functions studied in this paper will reveal closely related to a recently proposed voting rule, namely the k-wise Kemeny rule [10]. To compute a consensus rankings for the k-wise Kemeny rule, one needs only the information from the *choice matrix* derived from $\mathcal{P}$, denoted by $\mathcal{M}_{\mathcal{P}}$. The choice matrix gives, for each subset S of candidates and each candidate

c, the number of voters for which c is the preferred candidate in S. If only subsets of size at most k matter, the choice matrix can be restricted to these subsets.

Example 2. The choice matrix synthesizing the results of all setwise contests for the choice functions of Table 1 is given in Table 2. The matrix reads as follows: for instance, considering the rightmost column, one sees that c_1 is the most preferred candidate in $\{c_1, c_2, c_3\}$ for 3 voters, c_2 is the most preferred candidate for 2 voters, and c_3 is the most preferred candidate for 5 voters.

S	$\{c_1, c_2\}$	$\{c_1, c_3\}$	$\{c_2, c_3\}$	$\{c_1, c_2, c_3\}$
c_1	6	5	–	3
c_2	4	–	5	2
c_3	–	5	5	5

Table 2. The choice matrix for the instance of Example 1.

We now formally describe the k-wise Kemeny rule. Given a ranking r and a subset $S \in \Delta_k$ of candidates, let $t_r(S) \in S$ denote the most preferred candidate in S for r (i.e., for each candidate $c \neq t_r(S) \in S$, $t_r(S)$ is ranked at a higher position in r than c – is preferred to c). The k-wise distance $\delta^k(r, f)$ between a ranking r and a choice function $f \in \mathcal{F}_k$ is the number of disagreements between r and f on sets of candidates of size between 2 and k:

$$\delta^k(r, f) = \sum_{S \in \Delta_k} \mathbb{1}_{t_r(S) \neq f(S)}$$

where $\mathbb{1}_{t_r(S) \neq f(S)} = 1$ if $t_r(S) \neq f(S)$, 0 otherwise. Note that when $k = 2$, $\delta^2(r, f)$ is the well-known Kendall tau distance between r and $f \in \mathcal{F}_2$ (which associates a winner to each pair of candidates). We may also express δ^k by splitting Δ_k into sets of subsets of the same cardinality. Let us denote by $\mathcal{C}_i$ the set of subsets of $\mathcal{C}$ of cardinality equal to i. We have thus $\bigcup_{i=2}^{k} \mathcal{C}_i = \Delta_k$ and δ^k can be written:

$$\delta^k(r, f) = \sum_{i=2}^{k} \sum_{S \in \mathcal{C}_i} \mathbb{1}_{t_r(S) \neq f(S)}$$

Given a profile $\mathcal{P}$, the cost of a ranking r is the sum of the k-wise distances between r and each choice function f_j ($j \in \{1, \ldots, n\}$) in the choice function profile. It is thus the total number of disagreements between r and the voters on all the possible subsets of candidates of size at most k:

$$\delta^k(r, \mathcal{P}) = \sum_{j=1}^{n} \delta^k(r, f_j) = \sum_{j=1}^{n} \sum_{i=2}^{k} \sum_{S \in \mathcal{C}_i} \mathbb{1}_{t_r(S) \neq f_j(S)}$$

The k-wise Kemeny rule determines a ranking minimizing $\delta^k(r, \mathcal{P})$ among all the rankings $r \in R$. To compute such a consensus ranking, one needs only the information from the choice matrix $\mathcal{M}_{\mathcal{P}}$. It indeed minimizes $\delta^k(r, \mathcal{P})$:

$$\delta^k(r, \mathcal{P}) = \sum_{i=2}^{k} \sum_{S \in \mathcal{C}_i} d(\mathcal{M}_{\mathcal{P}}, S, r) \qquad (1)$$

where $d(\mathcal{M}_{\mathcal{P}}, S, r)$ is the number of voters whose most preferred candidate in S is *not* $t_r(S)$ (number of disagreements between r and $\mathcal{M}_{\mathcal{P}}$ for S). That is, a consensus ranking for the k-wise Kemeny rule minimizes the number of disagreements with the agents' choice functions on subsets of cardinality at most k (a disagreement occurs between a ranking r and choice function f on a subset S of candidates if $f(S) \neq t_r(S)$). The k-wise Kemeny rule generalizes the Kemeny rule, as it amounts to the usual Kemeny rule if $k=2$. Increasing k allows to overcome a well-known drawback of the Kemeny rule, namely that very different consensus rankings may coexist. The example below illustrates this.

Example 3. In Example 1, there are two consensus rankings for the Kemeny rule, namely $c_1 \succ c_2 \succ c_3$ and $c_3 \succ c_1 \succ c_2$ since both of them induce 14 pairwise disagreements with the preference profile. Candidate c_3 is thus ranked last in the former, and first in the latter. In contrast, for the 3-wise Kemeny rule, the only consensus ranking is $c_3 \succ c_1 \succ c_2$, with $14+5=19$ disagreements (14 on pairs, and 5 on $\{c_1, c_2, c_3\}$), while there are $14+7=21$ disagreements for $c_1 \succ c_2 \succ c_3$.

4 A Non-Utilitarian Discrete Choice Model

We now present the statistical model on choice functions that we will study in the remainder of the paper. The sample space (i.e., the possible observed outcomes from which the parameter of the statistical model are inferred) is the set of choice matrices. In this framework, the assumptions made by Condorcet and Young (see the introduction) need to be adapted, as we consider not only choices on pairs of candidates but also on subsets. Given a true ranking r^* of $\mathcal{C}$, the following assumptions are made on random variables $f_j(S)$ for all voters j:

1. for every $i \in \{2, \ldots, k\}$, $S \in \mathcal{C}_i$, $c \in S \setminus \{t_{r^*}(S)\}$, the probability that $f_j(S) = t_{r^*}(S)$ is $\alpha_i > 1$ times greater than the probability that $f_j(S) = c$: $Pr(f_j(S) = t_{r^*}(S)) = \alpha_i \cdot Pr(f_j(S) = c)$; that is, it is $\alpha_{|S|}$ more likely to choose the highest ranked candidate of S in r^* than any other given candidate of S.
2. for every pair $\{S, S'\}$ of subsets in Δ_k, $f_j(S)$ and $f_j(S')$ are independent.

For any pair $\{j, j'\}$ of voters, we also assume that each voter's preferred choice on each subset of candidates is independent of the other voters' preferences, i.e.:

3. for every $\{j, j'\} \subseteq \mathcal{V}$ and $(S, S') \in \Delta_k^2$, $f_j(S)$ and $f_{j'}(S')$ are independent.

For $k=2$, these assumptions amount to those made by Young on pairwise judgments in his analysis of Condorcet's theory of voting. If $k>2$, the additional parameters $\alpha_{|S|}$ (for $|S|>2$) give more flexibility to fit the observed choice data, at the cost of a greater computational load. Note that Assumption 1 means that the probability that voter j agree with ranking r^* on the preferred candidate in S depends only on the size of S, and not on the members of S.

Assumptions 1 and 2 yield the following statistical model for choice functions f, that we call *k-wise Young's model*, parameterized by a ranking r and choice probabilities $p_i = \alpha_i/(\alpha_i + i - 1)$ (conversely, $\alpha_i = (i-1)p_i/(1-p_i)$), where p_i represents $Pr(f(S) = t_r(S))$ for $|S| = i$:

Definition 1 (k-wise Young's Model). *Given a set $\mathcal{C}$ of m alternatives, the k-wise Young's model is defined as follows:*

- *the parameter space is $R\times\Theta$, where R is the set of rankings on $\mathcal{C}$ and $\Theta=(\frac{1}{2},1]\times\ldots\times(\frac{1}{k},1]$ is the set of choice probabilities $\overrightarrow{p}=(p_2,\ldots,p_k)$,*
- *for any $(r,\overrightarrow{p})\in R\times\Theta$, the probability $Pr(f|r,\overrightarrow{p})$ is*

$$\prod_{i=2}^{k}\prod_{S\in\mathcal{C}_i} p_i^{1-\mathbb{1}_{t_r(S)\neq f(S)}}\left(\frac{1-p_i}{i-1}\right)^{\mathbb{1}_{t_r(S)\neq f(S)}}$$

where $\mathbb{1}_{t_r(S)\neq f(S)}=1$ if $t_r(S)\neq f(S)$, 0 otherwise.

If $r=r^*$, we have indeed $Pr(f(S)=c)=(1–p_i)/(i–1)$ for $c\neq t_r(S)$ by Assumption 1, and the products in the formula for $Pr(f|r,\overrightarrow{p})$ follow from Assumption 2.

As the preferences revealed by the choices may be cyclic, sampling a choice function according to this model can be decomposed into independent draws for each subset $S\subseteq\mathcal{C}$. Given a choice function profile $\mathcal{P}$ with n voters, if one assumes the functions in $\mathcal{P}$ are *independently* sampled (in line with Assumption 3) from a k-wise Young's model of parameters r and $\overrightarrow{p}$, the probability $Pr(\mathcal{M}_\mathcal{P}|r,\overrightarrow{p})$ follows a multinomial distribution:

$$\prod_{i=2}^{k}\prod_{S\in\mathcal{C}_i}\frac{n!}{\prod_{c\in S}n_c!}p_i^{n-d(\mathcal{M}_\mathcal{P},S,r)}\left(\frac{1-p_i}{i-1}\right)^{d(\mathcal{M}_\mathcal{P},S,r)} \quad (2)$$

where n_c denotes the number of voters that choose candidate c in subset S.

From Equation 2, it is clear that the likelihood of $(r,\overrightarrow{p})$ given $\mathcal{M}_\mathcal{P}$, denoted by $\mathcal{L}(r,\overrightarrow{p}|\mathcal{M}_\mathcal{P})$, is proportional to

$$\prod_{i=2}^{k}\prod_{S\in C_i}p_i^{n-d(\mathcal{M}_\mathcal{P},S,r)}\left(\frac{1-p_i}{i-1}\right)^{d(\mathcal{M}_\mathcal{P},S,r)} \quad (3)$$

because the coefficients $n!/(\prod_{c\in S}n_c!)$ depend neither on r nor on $\overrightarrow{p}$. Let us now study different voting rules arising from Eq. 3. Depending on whether or not restrictive assumptions are made about probabilities p_i, we show that a maximum likelihood ranking r is a consensus ranking for the k-wise Kemeny rule, or for a weighted variant whose parameters vary with $\mathcal{M}_\mathcal{P}$ and r.

5 MLE of the Parameters of the k-Wise Young's Model

A consensus ranking for the k-wise Kemeny rule is an MLE of a ground truth ranking r^* if one assumes that the choice function profile is sampled according to the k-wise Young's model when $\alpha_2=\ldots=\alpha_k=\alpha>1$, i.e., in a subset S, candidate $t_{r^*}(S)$ is the most likely to be chosen, with a probability α times greater than any other given member of S, whatever the size of S. More formally:

Proposition 1 *If there exists $\alpha>1$ such that $\alpha_2=\ldots=\alpha_k=\alpha$, then, given a choice matrix $\mathcal{M}_\mathcal{P}$, a ranking r has maximum likelihood for the k-wise Young's model iff it minimizes $\delta^k(r,\mathcal{P})$, i.e., ranking r is a consensus ranking for the k-wise Kemeny rule.*

Proof. Maximizing Equation 3 amounts to maximizing:

$$\log\left(\prod_{i=2}^{k}\prod_{S\in\mathcal{C}_i} p_i^{n-d(\mathcal{M}_{\mathcal{P}},S,r)}\cdot\left(\frac{1-p_i}{i-1}\right)^{d(\mathcal{M}_{\mathcal{P}},S,r)}\right)$$

$$=\sum_{i=2}^{k}\sum_{S\in\mathcal{C}_i}\left((n-d(\mathcal{M}_{\mathcal{P}},S,r))\cdot\log p_i + d(\mathcal{M}_{\mathcal{P}},S,r)\cdot\log\left(\frac{1-p_i}{i-1}\right)\right)$$

$$=\sum_{i=2}^{k}\sum_{S\in\mathcal{C}_i} n\log p_i - \sum_{i=2}^{k}\sum_{S\in\mathcal{C}_i} d(\mathcal{M}_{\mathcal{P}},S,r)\cdot\log\left(\frac{p_i}{\frac{1-p_i}{i-1}}\right)$$

For a given set of values p_i, determining a ranking r that maximizes the above formula is equivalent to minimizing:

$$\sum_{i=2}^{k}\sum_{S\in\mathcal{C}_i} d(\mathcal{M}_{\mathcal{P}},S,r)\cdot\log\left(\frac{p_i}{\frac{1-p_i}{i-1}}\right)$$

As p_i is the probability to choose $t_r(S)$ and $(1-p_i)/(i-1)$ the probability to choose any other member of S, we have $p_i/(1-p_i/(i-1))=\alpha_i$. Furthermore, by assumption, $\alpha_i=\alpha\ \forall i\in\{2,\ldots,k\}$. Consequently, the expression simplifies to:

$$(\log\alpha)\cdot\sum_{i=2}^{k}\sum_{S\in\mathcal{C}_i} d(\mathcal{M}_{\mathcal{P}},S,r)$$

The coefficient $\log\alpha$ is strictly positive because $\alpha>1$ by assumption, and it can therefore be omitted when minimizing according to r. From Equation 1, we have:

$$\sum_{i=2}^{k}\sum_{S\in\mathcal{C}_i} d(\mathcal{M}_{\mathcal{P}},S,r)=\delta^k(r,\mathcal{P})$$

Therefore, whatever the vector $\overrightarrow{p}$ of choice probabilities, a ranking r that maximizes $\mathcal{L}(r,\overrightarrow{p}|\mathcal{M}_{\mathcal{P}})$ minimizes $\delta^k(r,\mathcal{P})$, which concludes the proof. □

For $k=2$, this proposition amounts to the result of Young regarding the interpretation of the Kemeny rule as an MLE of a ground truth ranking.

If we do not assume that the α_i are equal, then the maximum likelihood ranking may depend on $\overrightarrow{\alpha}=(\alpha_2,\ldots,\alpha_k)$, and we need to determine[3] a *couple* $(r,\overrightarrow{\alpha})$ of maximum likelihood $\mathcal{L}(r,\overrightarrow{\alpha}|\mathcal{M}_{\mathcal{P}})$, even if we are only interested in r. Determining such a couple $(r,\overrightarrow{\alpha})$ defines a new voting rule in itself, which returns r as a consensus ranking. The following result shows that it can be formulated as a discrete optimization problem on the space of rankings, because, for each ranking r, there exists a closed-form expression to determine the corresponding maximum likelihood values α_i.

[3] From now on, we use indifferently $\overrightarrow{p}$ or $\overrightarrow{\alpha}$, because one vector can be inferred from the other.

Proposition 2 *Given a choice matrix $\mathcal{M}_\mathcal{P}$, a couple $(r, \overrightarrow{\alpha})$ has maximum likelihood for the k-wise Young's model if and only if ranking r minimizes*

$$\delta_\alpha^k(r,\mathcal{P}) - \sum_{i=2}^{k} \sum_{S\in\mathcal{C}_i} n \log \frac{\alpha_i}{\alpha_i + i - 1}, \tag{4}$$

$$\textit{where} \quad \delta_\alpha^k(r,\mathcal{P}) = \sum_{i=2}^{k} (\log \alpha_i) \sum_{S\in\mathcal{C}_i} d(\mathcal{M}_\mathcal{P}, S, r) \tag{5}$$

$$\textit{and} \quad \alpha_i = ((i-1)\cdot \sum_{S\in\mathcal{C}_i} (n - d(\mathcal{M}_\mathcal{P}, S, r)))/(\sum_{S\in\mathcal{C}_i} d(\mathcal{M}_\mathcal{P}, S, r)). \tag{6}$$

Proof. From the proof of Proposition 1, we know that a couple $(r, \overrightarrow{p})$ has maximum likelihood iff, for a given choice matrix $\mathcal{M}_\mathcal{P}$ and ranking r, it maximizes:

$$f(\overrightarrow{p}) = \sum_{i=2}^{k} \sum_{S\in\mathcal{C}_i} n \log p_i - \sum_{i=2}^{k} \sum_{S\in\mathcal{C}_i} d(\mathcal{M}_\mathcal{P}, S, r)\cdot \log\left(\frac{p_i}{\frac{1-p_i}{i-1}}\right) \tag{7}$$

To determine an optimum of function f, each component p_i can be optimized independently from the others, because each one appears in a different term of the sum from $i=2$ to k. Noting that $\sum_{C\in\mathcal{C}_i} n = \binom{m}{i}\cdot n$ as there are $\binom{m}{i}$ different subsets $C\in\mathcal{C}_i$, the partial derivative of order 1 is written as:

$$\frac{\partial f}{\partial p_i}(\overrightarrow{p}) = \frac{\binom{m}{i}\cdot n - \sum\limits_{S\in\mathcal{C}_i} d(\mathcal{M}_\mathcal{P}, S, r)}{p_i} - \frac{\sum\limits_{S\in\mathcal{C}_i} d(\mathcal{M}_\mathcal{P}, S, r)}{(1-p_i)}.$$

For $p_i \in [0,1]$, the derivative vanishes for:

$$p_i = \left(\binom{m}{i}\cdot n - \sum_{S\in\mathcal{C}_i} d(\mathcal{M}_\mathcal{P}, S, r)\right) / \left(\binom{m}{i}\cdot n\right).$$

It is easy to prove that $\frac{\partial^2 f}{\partial p_i^2}(\overrightarrow{p}) < 0$ for $p_i \in [0,1]$, thus the corresponding stationary point of f is a maximum. From the values p_i of maximum likelihood we derive the values α_i of maximum likelihood:

$$\alpha_i = \frac{p_i}{\frac{1-p_i}{(i-1)}} = (i-1)\cdot \frac{\sum_{S\in\mathcal{C}_i}(n - d(\mathcal{M}_\mathcal{P}, S, r))}{\sum_{S\in\mathcal{C}_i} d(\mathcal{M}_\mathcal{P}, S, r)}$$

The result is obtained by expressing Eq. 7 in function of α_i instead of p_i, and turning the maximization into a minimization of the opposite expression. □

Note that, according to Proposition 2, the maximum likelihood value of each p_i given r corresponds to the observed proportion of agreements between r and $\mathcal{P}$ on subsets of size i, which is consistent with intuition. The formula of the likelihood of a couple $(r, \overrightarrow{\alpha})$ is written as the sum of two terms:

- the term $\delta_\alpha^k(r,\mathcal{P})$ is a weighted sum of disagreements between r and $\mathcal{P}$, where the disagreements on subsets of size i are weighted by $\log \alpha_i$;

– the term $-\sum_{i=2}^{k}\sum_{S\in\mathcal{C}_i} n\log(\alpha_i/(\alpha_i+i-1)) = -\log\prod_{j=1}^{n}\prod_{i=2}^{k}\prod_{S\in\mathcal{C}_i} p_i$; as $\prod_{j=1}^{n}\prod_{i=2}^{k}\prod_{S\in\mathcal{C}_i} p_i \leq 1$, the opposite of its logarithm is positive, and the term is all the greater as the empirical probability that the n choice functions in $\mathcal{P}$ coincide with r is low.

Let us now present algorithms (an exact one and a heuristic one) to compute a maximum likelihood couple $(r, \overrightarrow{\alpha})$ given a choice matrix $\mathcal{M}_{\mathcal{P}}$.

6 Algorithms for Determining an MLE

A brute force method for determining a couple $(r, \overrightarrow{\alpha})$ of maximum likelihood given $\mathcal{P}$ consists of computing a vector $\overrightarrow{\alpha}$ of maximum likelihood for each ranking r (thanks to Prop. 2), and, turning $\overrightarrow{\alpha}$ into $\overrightarrow{p}$, retaining the couple $(r, \overrightarrow{p})$ that maximizes Eq. 3.

A Faster Exact Algorithm. It is possible to improve this procedure by considering only a *subset* of rankings r on the candidates. We know indeed from Proposition 2 that, for any given $\overrightarrow{\alpha}$, the corresponding maximum likelihood ranking r minimizes $\delta_\alpha^k(r,\mathcal{P})$ (see Eq. 5). Minimizing $\delta_\alpha^k(r,\mathcal{P})$ can be seen as a multi-objective optimization problem, by associating to each r the vector:

$$\overrightarrow{d}_{\mathcal{P}}(r) = \left(\sum_{S\in\mathcal{C}_2} d(\mathcal{M}_{\mathcal{P}}, S, r), \ldots, \sum_{S\in\mathcal{C}_k} d(\mathcal{M}_{\mathcal{P}}, S, r)\right)$$

In multi-objective optimization problems, the goal is often to enumerate all the Pareto optimal solutions, i.e., in our setting, the rankings r such that there does not exist another ranking r' for which $\overrightarrow{d}_{\mathcal{P}}(r') \leq \overrightarrow{d}_{\mathcal{P}}(r)$, where $\overrightarrow{x} \leq \overrightarrow{y}$ if $\forall i \in \{2,\ldots,k\}\ x_i \leq y_i$, and $\exists i \in \{2,\ldots,k\}\ x_i < y_i$. A ranking r minimizing $\delta_\alpha^k(r,\mathcal{P})$ is obviously Pareto optimal. Such a ranking actually belongs to a more restricted set: the set of *supported* solutions, i.e., those that optimizes a weighted sum of the objectives [9]. The weight assigned to each objective i is here $\log\alpha_i$. An even more restricted set can be considered: the set of *extreme* rankings. A Pareto optimal ranking r is *extreme* if $\overrightarrow{d}_{\mathcal{P}}(r)$ is a vertex of the convex hull of $\{\overrightarrow{d}_{\mathcal{P}}(r) : r \in R\}$ in the $(k-1)$-dimensional objective space, where R is the set of all rankings. Indeed, it is well-known in multi-objective optimization that, for each supported ranking r', there exists an extreme ranking r such that $\delta_\alpha^k(r,\mathcal{P}) = \delta_\alpha^k(r',\mathcal{P})$.

A recent work presents a method for enumerating the extreme solutions in multi-objective optimization problems [19]. Based on such a method, we design an exact procedure for determining a maximum likelihood pair $(r, \overrightarrow{\alpha})$ by Prop. 2:

1. Determine all the extreme rankings by using Przybylski et al.'s method [19];
2. For each extreme ranking r, compute by Equation 6 the vector $\overrightarrow{\alpha}_r$ such that $(r, \overrightarrow{\alpha})$ has maximum likelihood;
3. Return a couple $(r, \overrightarrow{\alpha}_r)$ that minimizes Equation 4.

Although this procedure allows us to reduce the number of rankings we need to consider, there are still many of them, especially when the value of k increases. For this reason, we now propose a faster heuristic giving a very good approximation of an optimal couple $(r, \overrightarrow{\alpha})$.

A Heuristic Algorithm. Instead of considering all the extreme rankings, we propose an Iterative Optimization (IO) heuristic, which alternates two steps:

- α-step: compute an $\overrightarrow{\alpha}$ of maximum likelihood given r by Equation 6;
- r-step: compute an r of maximum likelihood given $\overrightarrow{\alpha}$ by minimizing $\delta^k_\alpha(r, \mathcal{P})$ (see Equation 5).

The minimization of $\delta^k_\alpha(r, \mathcal{P})$ is performed thanks to a weighted variant of the dynamic programming algorithm proposed by Gilbert et al. [10] for the k-wise Kemeny rule. The two steps are alternated until the same ranking is found in two consecutive r-steps. The complexity of the dynamic programming algorithm that computes a ranking r minimizing $\delta^k_\alpha(r, \mathcal{P})$ is $O(2^m m^2 n)$, thus the heuristic is not polynomial time (but much faster than the exact algorithm, as will be seen later). We launch the algorithm from a given vector $\overrightarrow{\alpha}$ for the r-step (in the numerical tests, we have set the values α_i corresponding to $p_i = 1/i + (i-1)/(10i)$).

7 Numerical Tests

We report here the results of several experiments[4] to test the performance of our heuristic and the fitness of the k-wise Young's (k-wise) model compared to that of the Plackett-Luce (PL) model on synthetic and real-world data.

Instances. The tests are carried out both on real data sets from the Preflib library [16], and on three types of synthetic instances. The first type of synthetic instances are *uniform instances*, in which the preferences of each voter is a random ranking in the set R of all permutations. The second type of instances, called *PL instances*, are preference profiles generated thanks to the PL model [14, 18]. The third type of instances, called *k-wise instances*, are choice matrices generated with our model. Given a ground truth ranking r^*, the choice function of a voter is generated as follows: for each subset S of size i, the voter chooses the winner in S w.r.t. r^* with probability p_i, and chooses any other candidate in S with probability $(1-p_i)/(i-1)$. We set $k\!=\!m$ in all tests.

Performance of the heuristic. In order to evaluate the performance of the IO heuristic, we compare the log-likelihood (LL) of the returned pair $(r, \overrightarrow{\alpha})$ with the one obtained by the exact method. Denoting by $OPT(I)$ the value of the LL of an optimal pair for a given instance I and by $IO(I)$ the value of the LL of the pair returned by the IO method, we calculate the ratio $q\!=\!IO(I)/OPT(I)$. For

[4] All algorithms have been implemented in C++, and the tests have been carried out on an Intel Core I5-8250 1.6GHz processor with 8GB of RAM.

all the real instances from the PrefLib library, and for all tested PL instances, the heuristic always returns an optimal pair $(r, \overrightarrow{\alpha})$. On uniform instances, the result is not always optimal, but it is very close to the optimal LL: the ratio q is above 0.9999 on average. The heuristic provides an excellent approximation of an optimal pair $(r, \overrightarrow{\alpha})$. Regardless of the type of instance, the IO method is much faster than the exact multi-objective algorithm. For example, for $m = k = 8$, the IO method takes 260 (resp. 185) seconds on average to return a solution for uniform (resp. PL) instances whereas the exact algorithm requires 30 000 (resp. 1000) seconds on average to determine an optimal pair for the same instances.

Model fitness. We now compare our model with the PL model in terms of fitness with real-world data. We use instances from the sushi dataset [11], in which 5000 voters give their ranking over 10 kinds of sushis. We randomly draw $n \in \{50, 100\}$ voters among the 5000. We apply the exact solution procedure proposed above, and compare the results with those of the PL model. The likelihood of a choice matrix w.r.t. the PL model for choice functions is written as follows: $\prod_{i=2}^{k} \prod_{S \in \mathcal{C}_i} \frac{n!}{\prod_{c \in S} n_c!} \prod_{c \in S} \left(\frac{u_c}{\sum_{d \in S} u_d} \right)^{n_c}$ where n_c denotes the number of voters choosing candidate c in S, and u_c the utility of c. To compare the fitness of the models, we use the Bayesian Information Criterion –BIC– [21]. Regarding the k-wise model, we consider the case of constant α_i's (model α) and the general case where the α_i may vary (model α_i). We compute the ratio $BIC(\mu)/BIC(PL)$ for $\mu \in \{\alpha, \alpha_i\}$. For 50 voters, the obtained ratio is 1.054 (resp. 1.047) for model α (resp. α_i). This improves to 1.052 (resp. 1.045) for 100 voters. This shows that for this dataset, the fitness of models α and α_i is close to the PL model, although the fitness of the latter is slightly better (by 5% at most).

Cross comparison. We now compare the PL model and the k-wise model on PL instances and k-wise instances. In both cases, we compute a correlation factor ρ between the returned ranking and the ground truth ranking used for generation. The factor ρ is the Kendall-Tau distance normalized between 0 and 1 – 0 indicates that the two rankings are identical while 1 means that they are opposite. Figure 1 shows the mean value of ρ in function of the level of correlation of the voters' preferences, for k-wise instances (left) and PL instances (right). For k-wise instances, the correlation between the choice functions is controlled by setting $p_i = 1/i + x(1 - 1/i)$ for $x \in [0, 1]$: all choice functions are equally likely and independent from the ground truth ranking for $x = 0$, while all choice functions are perfectly consistent with the ground truth ranking for $x = 1$. For PL instances, the correlation between the rankings is controlled by setting $u_p = 1 + (m - p)x$ as utility of the candidate in position p in the ground truth ranking: the higher x, the stronger the correlation. As one would expect, the MLE ranking for the k-wise (resp. PL) model is closer to the ground truth ranking on k-wise (resp. PL) instances. Interestingly, the k-wise model performs better on PL instances than the PL model on k-wise instances. When instances are correlated enough, the MLE ranking for both models always correspond to the ground truth.

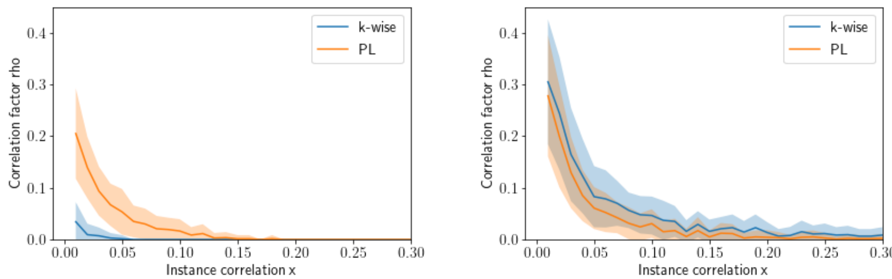

Fig. 1. Mean ρ (and 68% confidence interval) between the returned ranking and the ground truth on k-wise instances (left) and PL instances (right).

8 Conclusion

We have studied here an extension of Young's model for pairwise preferences to choices in subsets of size at most k, showing that the maximum likelihood ranking w.r.t. this model coincides with a consensus ranking for the k-wise Kemeny rule under certain assumptions on the choice probabilities. Relaxing these assumptions, we have proposed inference algorithms for the model, learning the choice probabilities from the data. The fitness of the model on real data is comparable with the Plackett-Luce model, although no utilities are embedded in our model. This is a first step towards the use of non-utilitarian discrete choice models for preference aggregation. For future work, correlating the choice probabilities with the ranks of the candidates within the considered subset (according to the ground truth ranking) is a natural research direction. Another direction is to investigate if the k-wise Young's model can be related to a k-wise distance-based statistical model for *rankings*, similarly to the connection between Young's model for pairwise preferences and Mallows' model [15] for rankings.

Acknowledgements. We acknowledge a financial support from the project THEMIS ANR20-CE23-0018 of the French National Research Agency (ANR).

References

1. Aleskerov, F.: Arrovian aggregation models. Kluwer Academic (1999)
2. Arrow, K.J.: Social choice and individual values (1951)
3. Baldiga, K.A., Green, J.R.: Assent-maximizing social choice. Social Choice and Welfare **40**(2), 439–460 (2013)
4. Caragiannis, I., Procaccia, A.D., Shah, N.: When do noisy votes reveal the truth? ACM Transactions on Economics and Computation (TEAC) **4**(3), 1–30 (2016)
5. Conitzer, V., Rognlie, M., Xia, L.: Preference functions that score rankings and maximum likelihood estimation. In: IJCAI, pp. 109–115 (2009)
6. Conitzer, V., Sandholm, T.: Common voting rules as maximum likelihood estimators. In: Proceedings of UAI 2005, pp. 145–152 (2005)
7. Cormack, G.V., Clarke, C.L., Buettcher, S.: Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: SIGIR, pp. 758–759 (2009)

8. Drissi-Bakhkhat, M., Truchon, M.: Maximum likelihood approach to vote aggregation with variable probabilities. Social Choice and Welfare **23**(2), 161–185 (2004)
9. Ehrgott, M.: Multicriteria optimization, vol. 491. Springer (2005)
10. Gilbert, H., Portoleau, T., Spanjaard, O.: Beyond pairwise comparisons in social choice: A setwise Kemeny aggregation problem. In: AAAI, pp. 1982–1989 (2020)
11. Kamishima, T.: Nantonac collaborative filtering: recommendation based on order responses. In: SIGKDD, pp. 583–588 (2003)
12. Kemeny, J.G.: Mathematics without numbers. Daedalus **88**(4), 577–591 (1959)
13. Lu, T., Boutilier, C.: The unavailable candidate model: a decision-theoretic view of social choice. In: Proceedings of EC 2010, pp. 263–274 (2010)
14. Luce, R.D.: Individual Choice Behavior: A Theoretical analysis. Wiley (1959)
15. Mallows, C.L.: Non-null ranking models. i. Biometrika **44**(1/2), 114–130 (1957)
16. Mattei, N., Walsh, T.: Preflib: A library of preference data HTTP://PREFLIB.ORG. In: ADT 2013, Lecture Notes in Artificial Intelligence, Springer (2013)
17. Pennock, D.M., Horvitz, E., Giles, C.L., et al.: Social choice theory and recommender systems: Analysis of the axiomatic foundations of collaborative filtering. In: AAAI/IAAI, pp. 729–734 (2000)
18. Plackett, R.L.: The analysis of permutations. Applied Stat. **24**(2), 193–202 (1975)
19. Przybylski, A., Klamroth, K., Lacour, R.: A simple and efficient dichotomic search algorithm for multi-objective mixed integer linear programs. arXiv (2019)
20. Raman, K., Joachims, T.: Methods for ordinal peer grading. In: SIGKDD, pp. 1037–1046 (2014)
21. Schwarz, G.: Estimating the dimension of a model. Ann. of stat. pp. 461–464 (1978)
22. Sen, A.: The possibility of social choice. American eco. rev. **89**(3), 349–378 (1999)
23. Soufiani, H.A., Parkes, D.C., Xia, L.: Random utility theory for social choice. In: NeurIPS, p. 126–134, NIPS'12, Curran Associates Inc. (2012)
24. Xia, L.: Learning and decision-making from rank data. Synthesis Lectures on Artificial Intelligence and Machine Learning **13**(1), 1–159 (2019)
25. Young, H.P.: Condorcet's theory of voting. American Political science review **82**(4), 1231–1244 (1988)