Ranking Median Regression: Learning to Order through Local Consensus

Stephan Clémençon, Anna Korba
LTCI Telecom ParisTech, Université Paris-Saclay
75013, Paris, France
{stephan.clemencon, anna.korba}@telecom-paristech.fr

Abstract

In the present era of personalized customer services and recommender systems, predicting the preferences of an individual/user over a (possibly very large) set of items indexed by $[\![n]\!] = \{1, \ldots, n\}$, $n \geq 1$, based on its characteristics, modelled as a r.v. X, taking its values in a feature space $\mathcal{X}$ say, is an ubiquitous issue. Though easy to state, this predictive problem is very difficult to solve in practice. The major challenge lies in the fact that, here, the (discrete) output space is the symmetric group $\mathfrak{S}_n$, composed of all permutations of $[\![n]\!]$, of explosive cardinality $n!$, and which is not a subset of a vector space. It is the purpose of this paper to explain how to build effectively nearly optimal predictive rules taking their values in $\mathfrak{S}_n$, by means of efficient ranking aggregation/consensus techniques implemented at a local level.

1 Introduction

In an increasing number of modern applications/interfaces, users are invited to declare their individual characteristics (*e.g.* socio-demographic features), taking the form of a random vector X valued in an input space $\mathcal{X} \subset \mathbb{R}^d$ say, and express their preferences on a collection of numbered services/products $[\![n]\!] = \{1, \ldots, n\}$ offered to them. In this context, the goal pursued is to learn from historical data how to predict the preferences of any user based on her characteristics X, the prediction being of the form of a permutation $s(X)$ on $[\![n]\!]$, mapping any item i to its rank $s(X)(i)$ on her preference list. Denoting by Σ the permutation that truly reflects the preferences of a user with characteristics X, the performance of any predictive rule, *i.e.* any measurable function $s : \mathcal{X} \to \mathfrak{S}_n$, can be measured by the expected Kendall τ distance between $s(X)$ and Σ

$$\mathcal{R}(S) = \mathbb{E}\left[d_\tau\left(s(X), \Sigma\right)\right], \tag{1}$$

where the expectation is taken over the (unknown) distribution of the pair (X, Σ) and $d_\tau(\sigma, \sigma') = \sum_{i<j} \mathbb{I}\{(\sigma(j) - \sigma(i)) \cdot (\sigma'(j) - \sigma'(i)) < 0\}$ for all $(\sigma, \sigma') \in \mathfrak{S}_n^2$, denoting by $\mathbb{I}\{\mathcal{E}\}$ the indicator function of any event $\mathcal{E}$. Stated this way, the objective is to build a mapping s that minimizes (1) and one may easily show with a straightforward conditioning argument that the optimal predictors are the rules that maps any point X in the input space to any (Kemeny) ranking median of P_X, Σ's conditional distribution given X. Recall that a Kemeny median of a probability distribution P on $\mathfrak{S}_n$ is any solution σ^* of the optimization problem

$$\min_{\sigma \in \mathfrak{S}_n} L_P(\sigma) \tag{2}$$

where $L_P(\sigma) = \mathbb{E}_{\Sigma \sim P}\left[d_\tau(\sigma, \Sigma)\right]$. For this reason, the predictive problem formulated above is referred to as *ranking median regression* in this paper. Here we investigate the latter (see also [1]) in the case where independent copies of (X, Σ) form the training dataset and show that techniques for statistical Kemeny aggregation, *i.e.* for solving (2), when implemented at local levels, may

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

produce efficient techniques to minimize (1) approximately (namely, nearest-neighbour and decision tree methods). Throughout the article, we denote by $L_P^* = L_P(\sigma^*)$ the minimum of (2) and by $\mathcal{R}^* = \mathbb{E}_{X\sim\mu}[L^*_{P_X}]$ the minimum of (1). The Dirac mass at any point a is denoted by δ_a.

2 Background and Preliminaries

2.1 A Statistical View of Consensus Ranking

Whereas problem (2) is NP-hard in general, exact solutions, referred to as *Kemeny medians*, can be explicited when the pairwise probabilities $p_{i,j} = \mathbb{P}\{\Sigma(i) < \Sigma(j)\}$, $1 \leq i \neq j \leq n$, fulfill the following property, referred to as *stochastic transitivity*.

Definition 1. *The probability distribution P on $\mathfrak{S}_n$ is stochastically transitive iff*

$$\forall (i,j,k) \in [\![n]\!]^3 : \;\; p_{i,j} \geq 1/2 \text{ and } p_{j,k} \geq 1/2 \;\Rightarrow\; p_{i,k} \geq 1/2.$$

If, in addition, $p_{i,j} \neq 1/2$ for all $i < j$, P is said to be strictly stochastically transitive.

When stochastic transitivity holds true, the set of Kemeny medians (see Theorem 5 in [2]) is the (non empty) set

$$\{\sigma \in \mathfrak{S}_n : \; (p_{i,j} - 1/2)(\sigma(j) - \sigma(i)) > 0 \text{ for all } i < j \text{ s.t. } p_{i,j} \neq 1/2\}, \tag{3}$$

and, if a strict version of stochastic transitivity is fulfilled (meaning that, in addition, none of the pairwise probabilities is equal to $1/2$), the Kemeny median is unique and given by the Copeland ranking:

$$\sigma_P^*(i) = 1 + \sum_{k \neq i} \mathbb{I}\{p_{i,k} < 1/2\} \text{ for } 1 \leq i \leq n. \tag{4}$$

We denote by $\mathcal{T}$ the set of strictly stochastically transitive distributions on $\mathfrak{S}_n$. Assume that we observe i.i.d. copies $\Sigma_1, \ldots, \Sigma_N$ of a generic r.v. $\Sigma \sim P$ and let $\widehat{P}_N = (1/N)\sum_{i=1}^N \delta_{\Sigma_i}$. Suppose that the underlying distribution P belongs to $\mathcal{T}$ and satisfies the low-noise condition $\mathbf{NA}(h)$ for a given $h > 0$:

$$\min_{i<j} |p_{i,j} - 1/2| \geq h. \tag{5}$$

It is shown in [2] that the empirical distribution $\widehat{P}_N$ is strictly stochastically transitive as well, with overwhelming probability, and that the expectation of the excess of risk of empirical Kemeny medians decays at an exponential rate, see Proposition 14 therein. In this case, the nearly optimal solution $\sigma^*_{\widehat{P}_N}$ can be made explicit and straightforwardly computed using Eq. (4) based on the empirical pairwise probabilities

$$\widehat{p}_{i,j} = \frac{1}{N}\sum_{k=1}^N \mathbb{I}\{\Sigma_k(i) < \Sigma_k(j)\}, \; i < j.$$

Otherwise, solving the NP-hard problem $\min_{\sigma\in\mathfrak{S}_n} L_{\widehat{P}_N}(\sigma)$ requires to get an empirical Kemeny median. However, as can be seen by examining Proposition 14's proof in [2], the exponential rate bound holds true for any candidate $\widetilde{\sigma}_N$ in $\mathfrak{S}_n$ that coincides with $\sigma^*_{\widehat{P}_N}$ when the empirical distribution lies in $\mathcal{T}$ and takes arbitrary values in $\mathfrak{S}_n$ otherwise. In practice, when $\widehat{P}_N$ does not belong to $\mathcal{T}$, we propose to consider as a pseudo-empirical median any permutation $\widetilde{\sigma}^*_{\widehat{P}_N}$ that ranks the objects as the empirical Borda count:

$$\left(\sum_{k=1}^N \Sigma_k(i) - \sum_{k=1}^N \Sigma_k(j)\right) \cdot \left(\widetilde{\sigma}^*_{\widehat{P}_N}(i) - \widetilde{\sigma}^*_{\widehat{P}_N}(j)\right) > 0 \text{ for all } i < j \text{ s.t. } \sum_{k=1}^N \Sigma_k(i) \neq \sum_{k=1}^N \Sigma_k(j),$$

breaking possible ties in an arbitrary fashion.

2.2 Statistical Framework for Ranking Median Regression

We assume now that we observe $(X_1, \; \Sigma_1) \; \ldots, \; (X_1, \; \Sigma_N)$ i.i.d. copies of the pair $(X, \; \Sigma)$ and, based on these training data, the objective is to build a predictive ranking rule s that nearly

minimizes $\mathcal{R}(s)$ over the class $\mathcal{S}$ of measurable mappings $s : \mathcal{X} \to \mathfrak{S}_n$. Of course, the Empirical Risk Minimization (ERM) paradigm encourages to consider solutions of the optimization problem:

$$\min_{s \in \mathcal{S}_0} \widehat{\mathcal{R}}_N(s), \tag{6}$$

where $\mathcal{S}_0$ is a subset of $\mathcal{S}$, supposed to be rich enough for containing approximate versions of elements of $\mathcal{S}^*$ (*i.e.* so that $\inf_{s \in \mathcal{S}_0} \mathcal{R}(s) - \mathcal{R}^*$ is 'small') and ideally appropriate for continuous or greedy optimization, and

$$\widehat{\mathcal{R}}_N(s) = \frac{1}{N} \sum_{k=1}^{N} d_\tau(s(X_k),\ \Sigma_k) \tag{7}$$

is a statistical version of (1) based on the (X_i, Σ_i)'s. Extending those established by [2] in the context of ranking aggregation, statistical results describing the generalization capacity of minimizers of (7) can be established under classic complexity assumptions for the class $\mathcal{S}_0$, as revealed by the result stated below.

Proposition 1. *Suppose that, for all $i < j$, the collection of sets*

$$\{\{x \in \mathcal{X} :\ s(x)(i) - s(x)(j) > 0\} :\ s \in \mathcal{S}_0\} \cup \{\{x \in \mathcal{X} :\ s(x)(i) - s(x)(j) < 0\} :\ s \in \mathcal{S}_0\}$$

is of finite VC *dimension $V < \infty$. Let $\widehat{s}_N$ be any minimizer of the empirical risk* (7) *over $\mathcal{S}_0$. For any $\delta \in (0,1)$, we have with probability at least $1 - \delta$: $\forall N \geq 1$,*

$$\mathcal{R}(\widehat{s}_N) - \mathcal{R}^* \leq C\sqrt{\frac{V \log(n(n-1)/(2\delta))}{N}} + \left\{\mathcal{R}^* - \inf_{s \in \mathcal{S}_0} \mathcal{R}(s)\right\}, \tag{8}$$

where $C < +\infty$ is a universal constant.

One may also prove that rates of convergence for the excess of risk of empirical Kemeny medians can be much faster than $O_{\mathbb{P}}(1/\sqrt{N})$ under the following hypothesis (generalizing (5)), involved in the subsequent analysis ([1]).

Assumption 1. *For all $x \in \mathcal{X}$, $P_x \in \mathcal{T}$ and $H = \inf_{x \in \mathcal{X}} \min_{i<j} |p_{i,j}(x) - 1/2| > 0$.*

This condition is checked in many situations, including most conditional parametric models (see Remark 13 in [2]), and generalizes condition (5), which corresponds to Assumption 1 when X and Σ are independent. The result stated below reveals that a similar fast rate phenomenon occurs for minimizers of the empirical risk (7) if Assumption 1 is satisfied.

Proposition 2. *Suppose that Assumption 1 is fulfilled, that the cardinality of class $\mathcal{S}_0$ is equal to $C < +\infty$ and that the unique true risk minimizer $s^*(x) = \sigma^*_{P_x}$ belongs to $\mathcal{S}_0$. Let $\widehat{s}_N$ be any minimizer of the empirical risk* (7) *over $\mathcal{S}_0$. For any $\delta \in (0,1)$, we have with probability at least $1 - \delta$:*

$$\mathcal{R}(\widehat{s}_N) - \mathcal{R}^* \leq \left(\frac{n(n-1)}{2H}\right) \times \frac{\log(C/\delta)}{N}. \tag{9}$$

3 Local Consensus Methods for Ranking Median Regression

We start here with introducing notations to describe the class of piecewise constant ranking rules and explore next approximation of a given ranking rule $s(x)$ by elements of this class, based on a local version of the concept of Kemeny median. Two strategies are next investigated in order to generate adaptively a partition tailored to the training data and yielding a ranking rule with nearly minimum predictive error. Throughout this section, for any measurable set $\mathcal{C} \subset \mathcal{X}$ weighted by $\mu(x)$, the conditional distribution of Σ given $X \in \mathcal{C}$ is denoted by $P_{\mathcal{C}}$. When it belongs to $\mathcal{T}$, the unique median of distribution $P_{\mathcal{C}}$ is denoted by $\sigma^*_{\mathcal{C}}$ and referred to as the local median on region $\mathcal{C}$.

3.1 Piecewise Constant Predictive Ranking Rules

Let $\mathcal{P}$ be a partition of $\mathcal{X}$ composed of $K \geq 1$ cells $\mathcal{C}_1, \ \ldots, \ \mathcal{C}_K$ (*i.e.* the $\mathcal{C}_k$'s are pairwise disjoint and their union is the whole feature space $\mathcal{X}$). Suppose in addition that $\mu(\mathcal{C}_k) > 0$ for

$k = 1, \ldots, K$. Using the natural embedding $\mathfrak{S}_n \subset \mathbb{R}^n$, any ranking rule $s \in \mathcal{S}$ that is constant on each subset $\mathcal{C}_k$ can be written as

$$s_{\mathcal{P},\bar{\sigma}}(x) = \sum_{k=1}^{K} \sigma_k \cdot \mathbb{I}\{x \in \mathcal{C}_k\}, \tag{10}$$

where $\bar{\sigma} = (\sigma_1, \ldots, \sigma_K)$ is a collection of K permutations. We denote by $\mathcal{S}_{\mathcal{P}}$ the collection of all ranking rules that are constant on each cell of $\mathcal{P}$. The following result describes the most accurate ranking median regression function in this class.

Proposition 3. *If $P_{\mathcal{C}_k} \in \mathcal{T}$ for $1 \leq k \leq K$, there exists a unique minimizer given by: $\forall x \in \mathcal{X}$,*

$$s^*_{\mathcal{P}}(x) = \sum_{k=1}^{K} \sigma^*_{P_{\mathcal{C}_k}} \cdot \mathbb{I}\{x \in \mathcal{C}_k\}. \tag{11}$$

We now investigate to what extent ranking median regression functions $s^*(x)$ can be well approximated by predictive rules of the form (10).

Assumption 2. *For all $1 \leq i < j \leq n$, the mapping $x \in \mathcal{X} \mapsto p_{i,j}(x)$ is Lipschitz, i.e. there exists $M < \infty$ such that:*

$$\forall (x, x') \in \mathcal{X}^2, \ \sum_{i<j} |p_{i,j}(x) - p_{i,j}(x')| \leq M \cdot ||x - x'||. \tag{12}$$

The following result shows that, under the assumptions above, the optimal prediction rule $\sigma^*_{P_X}$ can be accurately approximated by (11), provided that the regions $\mathcal{C}_k$ are 'small' enough.

Theorem 1. *Suppose that Assumptions 1-2 are fulfilled and that $P_{\mathcal{C}} \in \mathcal{T}$ for all $\mathcal{C} \in \mathcal{P}$. Then, we have:*

$$\mathbb{E}\left[d_\tau\left(\sigma^*_{P_X}, s^*_{\mathcal{P}}(X)\right)\right] \leq \sup_{x \in \mathcal{X}} d_\tau\left(\sigma^*_{P_x}, s^*_{\mathcal{P}}(x)\right) \leq (M/H) \cdot \delta_{\mathcal{P}}, \tag{13}$$

where $\delta_{\mathcal{P}} = \max_{\mathcal{C} \in \mathcal{P}} \sup_{(x,x') \in \mathcal{C}^2} ||x - x'||$ is the maximal diameter of $\mathcal{P}$'s cells. Hence, if $(\mathcal{P}_m)_{m \geq 1}$ is a sequence of partitions of $\mathcal{X}$ such that $P_{\mathcal{C}} \in \mathcal{T}$ for all $\mathcal{C} \in \mathcal{P}_m$, $m \geq 1$, and $\delta_{\mathcal{P}_m} \to 0$ as m tends to infinity, then

$$\sup_{x \in \mathcal{X}} d_\tau\left(\sigma^*_{P_x}, s^*_{\mathcal{P}_m}(x)\right) \to 0, \text{ as } m \to \infty.$$

3.2 Algorithms and Results

Nearest-Neighbor Rules for Ranking Median Regression. Fix $k \in \{1, \ldots, N\}$ and a query point $x \in \mathcal{X}$. Sort the training data $(X_1, \Sigma_1), \ldots, (X_n, \Sigma_n)$ by increasing order of the distance to x, measured, for simplicity, by $\|X_i - x\|$ for a certain norm chosen on $\mathcal{X} \subset \mathbb{R}^d$ say: $\|X_{(1,N)} - x\| \leq \ldots \leq \|X_{(N,N)} - x\|$. Consider next the empirical distribution calculated using the k training points closest to x

$$\widehat{P}(x) = \frac{1}{k} \sum_{l=1}^{k} \delta_{\Sigma_{(l,N)}} \tag{14}$$

and compute next the (pseudo)-empirical Kemeny median, as described in subsection 2.1, yielding the k-NN prediction at x:

$$s_{k,N}(x) \stackrel{def}{=} \widetilde{\sigma}^*_{\widehat{P}(x)}. \tag{15}$$

Observe incidentally that, under Assumptions 1-2, if $\|X_{(k,N)} - x\| < H/M$ then $\widehat{P}$ is necessarily strictly stochastically transitive. The result stated below provides an upper bound for the expected risk excess of (15), which reflects the usual bias/variance trade-off ruled by k for fixed N and asymptotically vanishes as soon as $k \to \infty$ as $N \to \infty$ such that $k = o(N)$. Notice incidentally that the choice $k \sim N^{2/(d+2)}$ yields the asymptotically optimal upper bound, of order $N^{-1/(2+d)}$.

Theorem 2. *Suppose that Assumptions 1-2 are fulfilled, that the r.v. X is bounded and $d \geq 3$. Then, we have: $\forall N \geq 1$, $\forall k \in \{1, \ldots, N\}$,*

$$\mathbb{E}\left[\mathcal{R}(s_{k,N}) - \mathcal{R}^*\right] \leq \frac{n(n-1)}{2}\left(\frac{1}{\sqrt{k}} + 2\sqrt{c_1}M\left(\frac{k}{N}\right)^{1/d} + \frac{2\sqrt{c_2}M}{H}\frac{1}{(N-k)^{1/d}}\right) \tag{16}$$

where c_1 and c_2 are constant which only depend on μ's support.

The implementation of this simple local method for ranking median regression does not require to explicit the underlying partition but is classically confronted with the curse of dimensionality. The next subsection explains how another local method, based on the popular tree induction heuristic, scales with the dimension of the input space by contrast.

Recursive Partitioning (CRIT). We now describe an iterative scheme for building an appropriate tree-structured partition $\mathcal{P}$, adaptively from the training data. Whereas the splitting criterion in most recursive partitioning methods is heuristically motivated, the local learning method we describe below relies on the ERM principle formulated in subsection 2.2, so as to build by refinement a partition $\mathcal{P}$ based on a training sample $\mathcal{D}_N = \{(\Sigma_1,\ X_1),\ \ldots,\ (\Sigma_N,\ X_N)\}$ so that, on each cell $\mathcal{C}$ of $\mathcal{P}$, the Σ_i's lying in it exhibit a small variability in the Kendall τ sense and, consequently, may be accurately approximated by a local Kemeny median. The goal pursued is thus to construct recursively a piecewise constant ranking rule associated to a partition $\mathcal{P}$, $s_{\mathcal{P}}(x) = \sum_{\mathcal{C}\in\mathcal{P}} \sigma_{\mathcal{C}} \cdot \mathbb{I}\{x \in \mathcal{C}\}$, with minimum empirical risk

$$\widehat{L}_N(s_{\mathcal{P}}) = \sum_{\mathcal{C}\in\mathcal{P}} \widehat{\mu}_N(\mathcal{C}) L_{\widehat{P}_{\mathcal{C}}}(\sigma_{\mathcal{C}}), \tag{17}$$

where $\widehat{\mu}_N = (1/N)\sum_{k=1}^N \delta_{X_k}$ is the empirical measure of the X_k's and, for any measurable subset $\mathcal{C} \subset \mathcal{X}$, $N_{\mathcal{C}} = \sum_{k=1}^N \mathbb{I}\{X_k \in \mathcal{C}\}$ and $\widehat{P}_{\mathcal{C}} = (1/N_{\mathcal{C}})\sum_{k:X_k\in\mathcal{C}} \delta_{\Sigma_k}$ is the empirical version of Σ's conditional distribution given $X \in \mathcal{C}$. We also denote by $\widehat{p}_{i,j}(\mathcal{C}) = (1/N_{\mathcal{C}})\sum_{k:\ X_k\in\mathcal{C}} \mathbb{I}\{\Sigma_k(i) < \Sigma_k(j)\}$, $i < j$, the local pairwise empirical probabilities. The partition $\mathcal{P}$ being fixed, the quantity (17) is minimum when $\sigma_{\mathcal{C}}$ is a Kemeny median of $\widehat{P}_{\mathcal{C}}$ for all $\mathcal{C} \in \mathcal{P}$. It is then equal to

$$\min_{s\in\mathcal{S}_{\mathcal{P}}} \widehat{L}_N(s) = \sum_{\mathcal{C}\in\mathcal{P}} \widehat{\mu}_N(\mathcal{C}) L^*_{\widehat{P}_{\mathcal{C}}}. \tag{18}$$

Except in the case where the intra-cell empirical distributions $\widehat{P}_{\mathcal{C}}$'s are all stochastically transitive, computing (18) at each recursion of the algorithm can be very expensive, since it involves the computation of a Kemeny median within each cell $\mathcal{C}$. We propose to measure instead the accuracy of the current partition by the quantity

$$\widehat{\gamma}_{\mathcal{P}} = \sum_{\mathcal{C}\in\mathcal{P}} \widehat{\mu}_N(\mathcal{C})\gamma_{\widehat{P}_{\mathcal{C}}},\ \text{where}\ \gamma_{\widehat{P}_{\mathcal{C}}} = \frac{1}{2}\sum_{i<j} \widehat{p}_{i,j}(\mathcal{C})\left(1 - \widehat{p}_{i,j}(\mathcal{C})\right)\ \text{for all}\ \mathcal{C} \in \mathcal{P}, \tag{19}$$

which satisfies the double inequality

$$\widehat{\gamma}_{\mathcal{P}} \leq \min_{s\in\mathcal{S}_{\mathcal{P}}} \widehat{L}_N(s) \leq 2\widehat{\gamma}_{\mathcal{P}}. \tag{20}$$

As shown above, the local variability measure we consider can be connected to the local ranking median regression risk and leads to exactly the same node impurity measure as in the tree induction method proposed in [3]. The algorithm described below differs from it in the method we use to compute the local predictions. The impurity of a cell $\mathcal{C}$ is thus measured by $\gamma_{\widehat{P}_{\mathcal{C}}}$ and a ranking median regression tree of maximal depth $J \geq 0$ is grown as follows. One starts from the root node $\mathcal{C}_{0,0} = \mathcal{X}$. At depth level $0 \leq j < J$, any cell $\mathcal{C}_{j,k}$, $0 \leq k < 2^j$ shall be split into two (disjoint) subsets $\mathcal{C}_{j+1,2k}$ and $\mathcal{C}_{j+1,2k+1}$, respectively identified as the left and right children of the interior leaf (j,k) of the ranking median regression tree, according to the following *splitting rule*.

Splitting rule. For any candidate left child $\mathcal{C} \subset \mathcal{C}_{j,k}$, picked in a class $\mathcal{G}$ of 'admissible' subsets (*e.g.* axis perpendicular splits), the relevance of the split $\mathcal{C}_{j,k} = \mathcal{C} \cup (\mathcal{C}_{j,k} \setminus \mathcal{C})$ is naturally evaluated through the quantity:

$$\Lambda_{j,k}(\mathcal{C}) \stackrel{def}{=} \widehat{\mu}_N(\mathcal{C})\gamma_{\widehat{P}_{\mathcal{C}}} + \widehat{\mu}_N(\mathcal{C}_{j,k} \setminus \mathcal{C})\gamma_{\widehat{P}_{\mathcal{C}_{j,k}\setminus\mathcal{C}}}. \tag{21}$$

Finding the best split thus consists in computing a solution $\mathcal{C}_{j+1,2k}$ of the optimization problem

$$\min_{\mathcal{C}\in\mathcal{G},\ \mathcal{C}\subset\mathcal{C}_{j,k}} \Lambda_{j,k}(\mathcal{C}), \tag{22}$$

which can be solved very efficiently, in a greedy fashion, when considering axis perpendicular splits for instance.

D_i	Setting 1			Setting 2			Setting 3		
	n=3	n=5	n=8	n=3	n=5	n=8	n=3	n=5	n=8
D_0	0.0698*	0.1290*	0.2670*	0.0173*	0.0405*	0.110*	0.0112*	0.0372*	0.0862*
	0.0473**	0.136**	0.324**	0.0568**	0.145**	0.2695**	0.099**	0.1331**	0.2188**
	(0.578)	(1.147)	(2.347)	(0.596)	(1.475)	(3.223)	(0.5012)	(1.104)	(2.332)
D_1	0.3475 *	0.569*	0.9405 *	0.306*	0.494*	0.784*	0.289*	0.457*	0.668*
	0.307**	0.529**	0.921**	0.308**	0.536**	0.862**	0.3374**	0.5714**	0.8544**
	(0.719)	(1.349)	(2.606)	(0.727)	(1.634)	(3.424)	(0.5254)	(1.138)	(2.287)
D_2	0.8656*	1.522*	2.503*	0.8305 *	1.447 *	2.359*	0.8105*	1.437*	2.189*
	0.7228**	1.322**	2.226**	0.723**	1.3305**	2.163**	0.7312**	1.3237**	2.252**
	(0.981)	(1.865)	(3.443)	(1.014)	(2.0945)	(4.086)	(0.8504)	(1.709)	(3.005)

Table 1: Empirical risk averaged on 50 trials on simulated data.

Local medians. The consensus ranking regression tree is grown until depth J and on each terminal leave $\mathcal{C}_{J,l}$, $0 \leq l < 2^J$, one computes the local Kemeny median estimate by means of the best strictly stochastically transitive approximation method investigated in subsection 2.1

$$\sigma^*_{J,l} \stackrel{def}{=} \widetilde{\sigma}^*_{\widehat{P}_{\mathcal{C}_{J,l}}}. \tag{23}$$

Experiments. We generated datasets of full rankings on n items according to two explanatory variables, varying the number of items ($n = 3, 5, 8$) and the nature of the features: in Setting 1, both features are numerical; in Setting 2, one is numerical and the other categorical, in Setting 3, both are categorical. For a fixed setting, a partition $\mathcal{P}$ of $\mathcal{X}$ composed of K cells $\mathcal{C}_1, \ldots, \mathcal{C}_K$ is fixed. In each trial, K permutations $\sigma_1, \ldots, \sigma_K$ (which can be arbitrarily close) are generated, as well as three datasets of N samples, where on each cell $\mathcal{C}_k$: the first one (denoted D_0) is constant (all samples are equal to σ_k), and the two others (denoted D_1 and D_2) are noisy versions of the first one, where the samples follow a Mallows distribution centered on σ_k with dispersion parameter $\phi = 2$ and $\phi = 1$ respectively. We choose K=6 and N=1000. The baseline model to which we compare our algorithms is the following: on the train set, we fit a K-means (with K=6), train a Plackett-Luce model on each cluster and assign the mode of this learnt distribution as the center ranking of the cluster. Results of the k-NN algorithm (indicated with a star *), of the CRIT algorithm (indicated with two stars **) and of the baseline model (between parenthesis) are provided Table 1. They show that the methods we develop recover well the underlying partition of the data.

4 Conclusions and Perspectives

The contribution of this article is twofold. The problem of learning to predict preferences, expressed in the form of a permutation, in a supervised setting is formulated and investigated in a rigorous probabilistic framework (optimal elements, learning rate bounds, bias analysis), extending that recently developped for statistical Kemeny ranking aggregation in [2]. Based on this formulation, it is also shown that predictive methods based on the concept of local Kemeny consensus, variants of nearest-neighbor and tree-induction methods namely, are well-suited for this learning task. This is justified by approximation theoretic arguments and algorithmic simplicity/efficiency both at the same time and illustrated by numerical experiments. Whereas the ranking median regression problem is motivated by many applications in our era of recommender systems and personalized customer services, the output variable may take the form of an *incomplete ranking* rather than a full ranking in many situations. Extension of our results to this more general framework, extended to incomplete rankings, will be the subject of further research.

References

[1] S. Clémençon, A. Korba, and E. Sibony. Ranking median regression: Learning to order through local consensus. In *Submitted*, 2017.

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

[3] P. L. H. Yu, W. M. Wan, and P. H. Lee. *Preference Learning*, chapter Decision tree modelling for ranking data, pages 83–106. Springer, New York, 2010.