Mach Learn (2013) 91:67–104
DOI 10.1007/s10994-012-5325-4

Ranking data with ordinal labels: optimality and pairwise aggregation

Stéphan Clémençon · Sylvain Robbiano · Nicolas Vayatis

Received: 8 March 2011 / Accepted: 24 September 2012 / Published online: 12 December 2012
© The Author(s) 2012

Abstract The paper describes key insights in order to grasp the nature of K-partite ranking. From the theoretical side, the various characterizations of optimal elements are fully described, as well as the *likelihood ratio monotonicity* condition on the underlying distribution which guarantees that such elements do exist. Then, a pairwise aggregation procedure based on Kendall tau is introduced to relate learning rules dedicated to bipartite ranking and solutions of the K-partite ranking problem. Criteria reflecting ranking performance under these conditions such as the ROC surface and its natural summary, the volume under the ROC surface (VUS), are then considered as targets for empirical optimization. The consistency of pairwise aggregation strategies are studied under these criteria and shown to be efficient under reasonable assumptions. Eventually, numerical results illustrate the relevance of the methodology proposed.

Keywords K-partite ranking · Ordinal data · ROC surface · Volume under the ROC surface · Empirical risk minimization · Median ranking

1 Introduction

In many situations, a natural ordering can be considered over a set of observations. When observations are documents in information retrieval applications, the ordering reflects degree of relevance for a specific query. In order to predict future ordering on new data, the learning process uses past data for which some relevance feedback is some provided,

Editor: Johannes Fürnkranz.

S. Clémençon · S. Robbiano (✉)
LTCI UMR Telecom ParisTech/CNRS No. 5141, Telecom ParisTech, Paris cedex 13, 75634, France
e-mail: robbiano@telecom-paristech.fr

S. Clémençon
e-mail: stephan.clemencon@telecom-paristech.fr

N. Vayatis
CMLA UMR CNRS No. 8536, ENS Cachan & UniverSud, Cahan cedex, 94235, France

such as ratings, say from 0 to 4, from the poorly relevant to the extremely relevant. For an example of such data, we refer to the LETOR benchmark data repository, see http://research.microsoft.com/en-us/um/people/letor/. A similar situation occurs in medical applications where decision-making support tools provide a scoring of the population of patients based on diagnostic test statistics in order to rank the individuals according to the advance state of a disease which are described as discrete grades, see Pepe (2003), Dreiseitl et al. (2000), Edwards et al. (2005), Mossman (1999) or Nakas and Yiannoutsos (2004) for instance.

A particular case which has received increasing attention both in machine learning the statistics literature is when only binary feedback is available (relevant *vs.* not relevant, ill *vs.* healthy) and this is known as the *bipartite ranking problem* (see Clémençon and Vayatis 2009b, 2010; Freund et al. 2003; Agarwal et al. 2005; Clémençon et al. 2008, *etc.*). In the presence of ordinal feedback (*i.e.* ordinal label taking a finite number of values, $K \geq 3$ say), the task consists in learning how to order temporarily unlabeled observations so as to reproduce as accurately as possible the ordering induced by the labels not observed yet. This problem is referred to as K-partite ranking and various approaches have been proposed in order to develop efficient algorithms in that case (see Rudin et al. 2005; Pahikkala et al. 2007). A closely related approach which points at both parametric and nonparametric statistical estimation is represented by *ordinal regression modeling* (see Waegeman et al. 2008b; Herbrich et al. 2000). To compare and assess the quality of these methods, a first concern is how to extend the typical performance measures such as the ROC curve and the AUC in this setup and this issue has been tackled in Scurfield (1996), Flach (2004). However, many interesting issues are still unexplored such as the theoretical optimality of learning rules, the statistical consistency of empirical performance maximization procedures, error bounds for K-partite ranking algorithms, ...

In the present paper, we tackle some of these open problems. In particular, we explore the connection between bipartite and K-partite ranking. Indeed, a natural approach is to transfer virtuous bipartite ranking methods to derive optimal and consistent rules for K-partite ranking. This idea is quite successful in the multiclass classification setup (see Hastie and Tibshirani 1998 or Fürnkranz 2002 for instance). We propose to build on the original proposition in Fürnkranz et al. (2009) to combine of bipartite ranking tasks in order to solve the K-partite case. A first intuition suggests that rules which are optimal for all bipartite ranking subproblems simultaneously should be optimal for the global problem. We offer examples in which this is not always the case and we state sufficient conditions for optimality which are called *monotonicity likelihood ratio* conditions. Based on this finding, we examine strategies which allow to combine rules dedicated for the pairwise subproblems for consecutive labels in order to derive interesting rules for the initial problem. We describe an efficient procedure for the pairwise aggregation of scoring rules which establishes a ranking consensus, called a *median scoring rule*, through an extension of the Kendall tau metric. It is also shown that such a median scoring rule always exists in the important situation where the scoring functions one seeks to summarize/aggregate are piecewise constant, and computation of this median rule is feasible. Next, we consider concepts such as the ROC surface and the Volume Under the ROC Surface (or VUS) which can be used to assess performance for scoring rules in K-partite ranking. Consistency can then be considered as convergence to optimal elements in terms of ROC surface or VUS. We then study conditions under which consistency of pairwise aggregation can be achieved. Indeed, it can be shown that under the monotone likelihood ratio condition together with a margin condition over the posterior distributions, the median scoring rule built out of pairwise AUC-consistent rules is VUS-consistent. We also consider specific strategies to derive scoring rules for this problem such

as the empirical maximization of the VUS or the plug-in scoring rule. We also provide an analysis of the empirical performance of the Kendall-type pairwise aggregation method using the TREERANK algorithm developed by the authors (Clémençon and Vayatis 2009b). An extensive comparison with state-of-the-art ranking methods is presented both on artificial and real data sets and we exhibit performance in terms of VUS, as well as the form of the level sets of the estimated scoring rules. The latter visualization show interesting insights about the geometry of risk segments in the input space.

The rest of the paper is structured as follows. In Sect. 2, the probabilistic setting is introduced and optimal scoring rules for K-partite ranking are successively defined and characterized. A specific *monotonicity likelihood ratio* condition is stated, which is shown to guarantee the existence of a natural optimal ordering over the input space. A novel Kendall-type aggregation procedure is presented in Sect. 3 and performance metrics, such as the VUS, are the subject matter of Sect. 4. Consistency results and insights on the passage from bipartite subproblems to the full K-partite case are discussed in Sect. 5. Finally, Sect. 6 displays a series of numerical results and illustrations for the aggregation principle considered in this paper. Mathematical proofs are postponed to the Appendix A.

2 Optimal elements in ranking data with ordinal labels

2.1 Probabilistic setup and notations

We consider a black-box system with random input/output pair (X, Y). We assume that the input random vector X takes values over $\mathbb{R}^d$ and the output Y over the ordered discrete set $\mathcal{Y} = \{1, \ldots, K\}$. Here it is assumed that the ordered values of the output Y can be reflected by an ordering over $\mathbb{R}^d$. The case where $K = 2$ is known as the bipartite ranking setup. In this paper, we focus on the case where $K > 2$. Though the objective pursued here is different, the probabilistic setup is exactly the same as that of *ordinal regression*, see Sect. 5.4 for a discussion of the connections between these two problems. We denote by f_k the density function of the class-conditional distributions of X given $Y = k$ and by $\mathcal{X}_k \subseteq \mathbb{R}^d$ the support of f_k. We also set $p_k = \mathbb{P}\{Y = k\}$, $k \in \{1, \ldots, K\}$, the mixture parameter for class $Y = k$, and $\eta_k(x) = \mathbb{P}\{Y = k \mid X = x\}$ the posterior probability. Set $f = p_1 f_1 + \cdots + p_K f_K$ and recall that:

$$\forall k \in \{1, \ldots, K\},\ \forall x \in \bigcup_{l=1}^{K} \mathcal{X}_l, \quad \eta_k(x) = p_k \cdot \frac{f_k}{f}(x).$$

The regression function $\eta(x) = \mathbb{E}[Y \mid X = x]$ can be expressed in the following way:

$$\forall x \in \bigcup_{l=1}^{K} \mathcal{X}_l, \quad \eta(x) = \sum_{k=1}^{K} k \cdot \eta_k(x),$$

as the expectation of a discrete random variable. We shall make use of the following notation for the likelihood ratio of the class-conditional distribution:

$$\forall k, l \in \{1, \ldots, K\},\ l < k,\ \forall x \in \mathcal{X}_l, \quad \Phi_{k,l}(x) = \frac{f_k}{f_l}(x) = \frac{p_k}{p_l} \cdot \frac{\eta_k}{\eta_l}(x).$$

Along the paper, we shall use the convention that $u/0 = \infty$ for any $u \in]0, \infty[$ and $0/0 = 0$.

2.2 Optimal scoring rules

The problem considered in this paper is to infer an order relationship over $\mathbb{R}^d$ after observing vector data with ordinal labels. For this purpose, we consider real-valued decision rules of the form $s : \mathbb{R}^d \to \mathbb{R}$ called *scoring rules*. In the case of ordinal labels, the main idea is that good scoring rules s are those which assign a high score $s(X)$ to the observations X with large values of the label Y. We now introduce the concept of optimal scoring rule for ranking data with ordinal labels.

Definition 1 (Optimal scoring rule) An optimal scoring rule s^* is a real-valued function such that:

$$\forall k, l \in \{1, \ldots, K\},\ l < k,\ \forall x, x' \in \mathcal{X}_l,\ \Phi_{k,l}(x) < \Phi_{k,l}\left(x'\right) \Rightarrow s^*(x) < s^*\left(x'\right).$$

The rationale behind this definition can be understood by considering the case $K = 2$. The class $Y = 2$ should receive higher scores than the class $Y = 1$. In this case, an optimal scoring rule s^* should score observations x in the same order as the posterior probability η_2 of the class $Y = 2$ (or equivalently as the ratio $\eta_2/(1 - \eta_2)$). Since $\eta_1(x) + \eta_2(x) = 1$, for all x, it is easy to see that this is equivalent to the condition described in the previous definition (see Clémençon and Vayatis 2009b for details). In the general case ($K > 2$), optimality of a scoring rule s^* means that s^* is optimal for all bipartite subproblems with classes $Y = k$ and $Y = l$, with $l < k$.

An important remark is that, in the probabilistic setup introduced above, an optimal scoring rule may not exist as shown in the next example.

Example 1 Consider a discrete input space $\mathcal{X} = \{x_1, x_2, x_3\}$ and $K = 3$. We assume the following joint probability distribution $\mathbb{P}(X = x_i, Y = j) = \omega_{i,j}$ for the random pair (X, Y):

$$\begin{aligned} \omega_{1,1} = \omega_{2,2} = \omega_{3,3} &= 1/2, \\ \omega_{1,2} = \omega_{2,3} = \omega_{3,1} &= 1/3, \\ \omega_{1,3} = \omega_{2,1} = \omega_{3,2} &= 1/6. \end{aligned}$$

Note that in the case of a discrete distribution for X, the density function coincides with mass function and we have $f(x) = \mathbb{P}(X = x)$. It is then easy to check that, in this case, there is no optimal scoring rule for this distribution in the sense of Definition 1.

2.3 Existence and characterization of optimal scoring rules

The previous example shows that the existence of optimal scoring rules cannot be guaranteed under any joint distribution. Our first important result is the characterization of those distributions for which the family of optimal scoring rules is not an empty set. The next proposition offers a necessary and sufficient condition on the distribution which ensures the existence of optimal scoring rules.

Assumption 1 For any $k, l \in \{1, \ldots, K\}$ such that $l < k$, for all $x, x' \in \mathcal{X}$, we have:

$$\Phi_{k+1,k}(x) < \Phi_{k+1,k}\left(x'\right) \Rightarrow \Phi_{l+1,l}(x) \leq \Phi_{l+1,l}\left(x'\right).$$

Proposition 1 *The following statements are equivalent*:

(1) *Assumption 1 holds*.
(2) *There exists an optimal scoring rule* s^*.
(3) *The regression function* $\eta(x) = \mathbb{E}(Y \mid X = x)$ *is an optimal scoring rule*.
(4) *For any* $k \in \{1, \ldots, K-1\}$, *for all* $x, x' \in \mathcal{X}_k$, *we have*:

$$\Phi_{k+1,k}(x) < \Phi_{k+1,k}\big(x'\big) \Rightarrow s^*(x) < s^*\big(x'\big).$$

(5) *For any* $k, l \in \{1, \ldots, K\}$ *such that* $l < k$, *the ratio* $\Phi_{k,l}(x)$ *is a nondecreasing function of* $s^*(x)$.

Assumption 1 characterizes the class distributions for the random pair (X, Y) for which the very concept of an optimal scoring rules makes sense. The proposition says that if this condition is not satisfied then the ordinal nature of the labels, when seen through the observation X, is violated. We point out that a related condition, called *ERA ranking representability*, has been introduced in Waegeman and Baets (2011), see Definition 2.1 therein. Precisely, it can be easily checked that the condition in the previous proposition means that the collection of (bipartite) ranking functions $\{\Phi_{k+1,k} : 1 \leq k < K\}$ is an ERA ranking representable set of ranking functions. Statement (3) suggests that plug-in rules based on the statistical estimation of the regression function η and multiple thresholding of the estimate will offer candidates for practical resolution of K-partite ranking. Such strategies are indeed reminiscent of ordinal logistic regression methods and will be discussed in Sect. 5.3.2. Statement (4) offers an alternative characterization to Definition 1 for optimal scoring rules. Statement (5) means that the family of densities of the class-conditional distributions f_k has a *monotone likelihood ratio* (we refer to standard textbooks of mathematical statistics which use this terminology, e.g. Lehmann and Romano 2005).

Proposition 2 *Under Assumption 1 we necessarily have*:

$$\mathcal{X}_{k'} \cap \mathcal{X}_{l'} \subseteq \mathcal{X}_k \cap \mathcal{X}_l \quad \text{for any } k, k', l, l' \text{ such that } 1 \leq k' \leq k < l \leq l' \leq K.$$

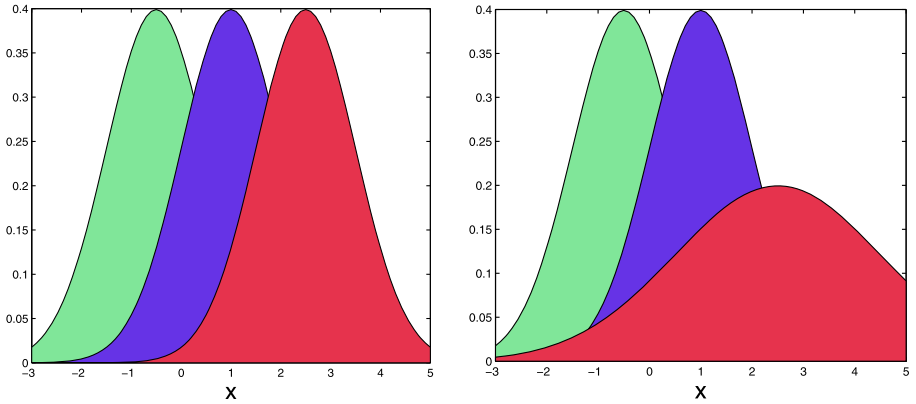
2.4 Examples and counterexamples of monotone likelihood ratio families

It is easy to see that, in absence of Assumption 1, the notion of K-partite ranking hardly makes sense. However, it is a challenging statistical task to assess whether data arise from a mixture of distributions F_k with monotone likelihood ratio. We now provide examples and counterexamples of such cases.

Disjoint supports Consider the separable case where: $\forall k, l$, $\mathcal{X}_k \cap \mathcal{X}_l = \emptyset$. Then Assumption 1 is clearly fulfilled as for $k \neq l$, we have either $\Phi_{k,l} = 0$ or ∞. It is worth mentioning that in this case, the nature of the K-partite ranking problem does not differ from the multi-class classification setup where there is no order relation between classes.

Exponential families We recall that $f = \sum_{k=1}^{K} p_k f_k$ is the marginal distribution function of X. We introduce the following choice for the class-conditional distributions f_k:

$$f_k(x) = \exp\big\{\kappa(k)T(x) - \psi(k)\big\} f(x), \quad \forall x \in \mathbb{R}^d,$$



(a) Gaussian class densities fulfilling Assumption 1 : $m_1 = -0.5,\ m_2 = 1,\ m_3 = 2.5,\ \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$

(b) Gaussian class densities not fulfilling Assumption 1 : $m_1 = -0.5,\ m_2 = 1,\ m_3 = 2.5,\ \sigma_1^2 = \sigma_2^2 = 1,\ \sigma_3^2 = 2$

Fig. 1 Two examples of 1-D conditional Gaussian distributions in the case $K = 3$—class 1 in *green*, class 2 in *blue* and class 3 in *red*

where:

- $\kappa : \{1, \ldots, K\} \to \mathbb{R}$ is strictly increasing,
- $T : \mathbb{R}^d \to \mathbb{R}$ such that $\psi(k) = \int_{x \in \mathbb{R}^d} \exp\{\kappa(k)T(x)\} f(x)dx < +\infty$, for $1 \leq k \leq K$.

It is easy to check that he family of density functions f_k has the property of monotone likelihood ratio.

1-*D Gaussian distributions* We consider here a toy example with $K = 3$ and the f_k are Gaussian distributions $\mathcal{N}(m_k, \sigma_k^2)$ over $\mathbb{R}$, where m_k is the expectation and σ_k^2 is the variance. Depending on the values of the parameters m_k, σ_k^2, the collection $\{f_1, f_2, f_3\}$ may or may not satisfy the property of having a monotone likelihood ratio. Assume first that the variances are equal, then the property of monotone likelihood ratio is satisfied if and only if either $m_1 \leq m_2 \leq m_3$ or $m_3 \leq m_2 \leq m_1$ (see Fig. 1(a)). Figure 1(b) depicts a situation where $m_1 < m_2 < m_3$ and $\sigma_3^2 > \sigma_2^2 = \sigma_1^2$ for which the random observation X does not permit to recover the preorder induced by the output variable. The monotonicity condition is violated for instance at $(x, x') = (-2, 1)$ and there is no optimal scoring rule in this case.

Uniform noise Let $t_0 = -\infty < t_1 < \cdots < t_{K-1} < +\infty$ and $g : \mathbb{R}^d \to \mathbb{R}$ a measurable function. We define the output random variable through:

$$Y = \sum_{k=1}^{K} \mathbb{I}\big\{g(X) + U > t_{k-1}\big\},$$

where U is a uniform random variable on some interval of the real line, independent from X. Then it can be easily seen that the class-conditional distributions form a collection with monotone likelihood ratio, provided that t_1 and t_{K-1} both lie inside the interval defined by the essential infimum and supremum of the random variable $g(X) + U$. In this case, any strictly increasing transform of g is an optimal scoring rule.

3 Pairwise aggregation: from bipartite to K-partite ranking

In the present section, we propose a practical strategy for building scoring rules which approximate optimal scoring rules for K-partite ranking based on data. The principle of this strategy is the aggregation of scoring rules obtained for the pairwise subproblems. We emphasize the fact that the situation is very different from multiclass classification where aggregation boils down to linear combination, or majority voting, over binary classifiers (for "*one against one*" and "*one versus all*", we refer to Allwein et al. 2001; Hastie and Tibshirani 1998; Venkatesan and Amit 1999; Debnath et al. 2004; Dietterich and Bakiri 1995; Beygelzimer et al. 2005a, 2005b and the references therein for instance). We propose here, in the K-partite ranking setup, a metric-based barycentric approach to build the aggregate scoring rule from the collection of scoring rules estimated for the bipartite subproblems. In order to avoid technical discussions dealing with special cases, we assume in the sequel that all class-conditional distributions have a continuous density f_k and share the same support $\mathcal{X} \subset \mathbb{R}^d$.

3.1 Median scoring rules and optimal aggregation

Every scoring rule induces an order relation over the input space $\mathbb{R}^d$ and, for the ranking problem considered here, a measure of similarity between two scoring functions should only take into consideration the similarity in the ranking induced by each one of them. We propose here a measure of agreement between scoring rules which is based on the probabilistic Kendall τ for a pair of random variables.

Definition 2 (Probabilistic Kendall τ) Consider X, X' i.i.d. random vectors with density function f over $\mathbb{R}^d$. The measure of agreement between two real-valued scoring rules s_1 and s_2 is defined as the quantity:

$$\begin{aligned}\tau(s_1, s_2) = {} & \mathbb{P}\big\{\big(s_1(X) - s_1\big(X'\big)\big) \cdot \big(s_2(X) - s_2\big(X'\big)\big) > 0\big\} \\ & + \frac{1}{2}\mathbb{P}\big\{s_1(X) \neq s_1\big(X'\big), s_2(X) = s_2\big(X'\big)\big\} \\ & + \frac{1}{2}\mathbb{P}\big\{s_1(X) = s_1\big(X'\big), s_2(X) \neq s_2\big(X'\big)\big\}.\end{aligned}$$

This definition of agreement between scoring rules s_1 and s_2 coincides indeed with the Kendall τ between real-valued random variables $s_1(X)$ and $s_2(X)$. Note that the contribution of the two last terms in the definition of $\tau(s_1, s_2)$ vanishes when the distributions of the $s_i(X)$'s are continuous.

From there one can define the notion of median scoring rule which accounts for the consensus of many real-valued scoring rules over a given class of candidates.

Definition 3 (Median scoring rule) Consider a given class $\mathcal{S}_1$ of real-valued scoring rules and $\Sigma_K = \{s_1, \ldots, s_{K-1}\}$ a finite set of real-valued scoring rules. A median scoring rule $\overline{s}$ for $(\mathcal{S}_1, \Sigma_K)$ satisfies:

$$\sum_{k=1}^{K-1} \tau(\overline{s}, s_k) = \sup_{s \in \mathcal{S}_1} \sum_{k=1}^{K-1} \tau(s, s_k). \tag{1}$$

In general, the supremum appearing on the right hand side of Eq. (1) is not attained. However, when the supremum over $\mathcal{S}_1$ can be replaced by a maximum over a finite set $\mathcal{S}_1' \subset \mathcal{S}_1$, a median scoring rule always exists (but it is not necessarily unique). In particular, this is the case when considering *piecewise constant scoring functions* such as those produced by the bipartite ranking algorithms proposed in Clémençon et al. (2011a), Clémençon and Vayatis (2009a, 2010) (we also refer to Clémençon and Vayatis 2009c for a discussion of consensus computation/approximation in this case). The idea underlying the measure of consensus through Kendall metric in order to aggregate scoring functions that are nearly optimal for bipartite ranking subproblems is clarified by the following result.

Definition 4 (PAIRWISE OPTIMAL SCORING RULE) A pairwise optimal scoring rule $s^*_{l,k}$ is an optimal scoring rule for the bipartite ranking problem with classes $Y = k$ and $Y = l$, where $k > l$ in the sense that:

$$\forall x, x' \in \mathcal{X},\ \Phi_{k,l}(x) < \Phi_{k,l}\left(x'\right) \Rightarrow s^*_{l,k}(x) < s^*_{l,k}\left(x'\right).$$

We denote by $\mathcal{S}^*_{l,k}$ the set of such optimal rules and, in particular, $\mathcal{S}^*_k = \mathcal{S}^*_{k,k+1}$.

Proposition 3 *Denote by $\mathcal{S}$ the set of all possible real-valued scoring rules and consider pairwise optimal scoring rules $s^*_k \in \mathcal{S}^*_k$ for $k = 1, \ldots, K-1$, which form the set $\Sigma^*_K = \{s^*_1, \ldots, s^*_{K-1}\}$. Under Assumption 1, we have*:

1. *A median scoring rule $\overline{s}^*$ for $(\mathcal{S}, \Sigma^*_K)$ is an optimal scoring rule for the K-partite ranking problem.*
2. *Any optimal scoring rule s^* for the K-partite ranking problem satisfies*:

$$\sum_{k=1}^{K-1} \tau\left(s^*, s^*_k\right) = K - 1.$$

The proposition above reveals that "consensus scoring rules", in the sense of Definition 3, based on $K-1$ optimal scoring rules are still optimal solutions for the global K-partite ranking problem and that, conversely, optimal elements necessarily achieve the equality in Statement (2) of the previous proposition. This naturally suggests to implement the following two-stage procedure, that consists in (1) solving the bipartite ranking subproblem related to the pairwise case $(k, k+1)$ of consecutive class labels, yielding a scoring function s_k, for $1 \leq k < K$, and (2) computing a median according to Definition 3, when feasible, based on the latter over a set $\mathcal{S}_1$ of scoring functions. Beyond the difficulty to solve each ranking subproblem separately (for instance refer to Clémençon and Vayatis 2009b for a discussion of the nature of the bipartite ranking issue), the performance/complexity of the method sketched above is ruled by the richness of the class $\mathcal{S}_1$ of scoring function candidates: too complex classes clearly make median computation unfeasible, while poor classes may not contain sufficiently accurate scoring rules.

3.2 A practical aggregation procedure

We now propose to convert the previous theoretical results which relate pairwise optimality to K-partite optimality in ranking into a practical aggregation procedure. Consider two independent samples:

- a sample $\mathcal{D} = \{(X_i, Y_i)\colon 1 \le i \le n\}$ with i.i.d. labeled observations,
- a sample $\mathcal{D}' = \{X'_i, : 1 \le i \le n'\}$ a sample with unlabeled observations.

The first sample $\mathcal{D}$ is used for training bipartite ranking rules $\hat{s}_k$, while the second sample $\mathcal{D}'$ will be used for the computation of the median. In practice a proxy for the median is computed based on the empirical version of the Kendall τ, the following U-statistic of degree two, see Clémençon et al. (2008).

Definition 5 (Empirical Kendall τ) Given a sample $X_1, \ldots, X_n$, the empirical Kendall τ is given by:

$$\widehat{\tau}_n(s_1, s_2) = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} h\big(\big(s_1(X_i), s_1(X_j)\big), \big(s_2(X_i), s_2(X_j)\big)\big),$$

where

$$h\big((v, w), \big(v', w'\big)\big) = \mathbb{I}\big\{\big(v - v'\big) \cdot \big(w - w'\big) > 0\big\} + \frac{1}{2}\mathbb{I}\big\{v = v', w \ne w'\big\} + \frac{1}{2}\mathbb{I}\big\{v \ne v', w = w'\big\},$$

for (v, w) and (v', w') in $\mathbb{R}^2$.

Denote by $\mathcal{D}_k = \{(X_i, Y_i) \in \mathcal{D}\colon Y_i = k\}$. The following aggregation method describes a two-steps procedure which takes as input the two data sets, a class $\mathcal{S}_1$ of candidate scoring rules, and a generic bipartite ranking algorithm $\mathcal{A}$.

Kendall aggregation for K-partite ranking

Input. Data samples $\mathcal{D}$ and $\mathcal{D}'$, a bipartite ranking algorithm $\mathcal{A}$, a class $\mathcal{S}_1$ of scoring rules.

1. **Build pairwise scoring rules for bipartite ranking.** For $k = 1, \ldots, K-1$, run algorithm $\mathcal{A}$ in order to train a scoring function $\widehat{s}_k(x)$ based on the restricted samples $\mathcal{D}_k \cup \mathcal{D}_{k+1}$.

2. **Aggregate pairwise scoring rules for K-partite ranking.** Compute

$$\widehat{s} = \arg\max_{s \in \mathcal{S}_1} \sum_{k=1}^{K-1} \widehat{\tau}'(s, \widehat{s}_k),$$

where $\widehat{\tau}'$ is the empirical Kendall τ computed over the sample $\mathcal{D}'$.

Output. Empirical median scoring rule $\widehat{s}$ in $\mathcal{S}_1$ for K-partite ranking

Practical implementation issues Motivated by practical problems such as the design of meta-search engines, collaborative filtering or combining results from multiple databases, *consensus ranking*, which the second stage of the procedure described above is a special case of, has recently enjoyed renewed popularity and received much attention in the

machine-learning literature, see Meila et al. (2007), Fagin et al. (2004) or Lebanon and Lafferty (2002) for instance. As shown in Hudry (2008) or Wakabayashi (1998) in particular, median computations are NP-hard problems in general. Except in the case where $\mathcal{S}_1$ is of very low cardinality, the (approximate) computation of a supremum involves in practice the use of meta-heuristics such as simulated annealing, tabu search or genetic algorithms. The description of these computational approaches to consensus ranking is beyond the scope of this paper and we refer to Barthélemy et al. (1989), Charon and Hudry (1998), Laguna et al. (1999) or Mandhani and Meila (2009) and the references therein for further details on their implementation. We also underline that the implementation of the Kendall aggregation approach could be naturally based on $K(K-1)/2$ scoring functions, corresponding to solutions of the bipartite subproblems defined by all possible pairs of labels (the theoretical analysis carried out below can be straightforwardly extended so as to establish the validity of this variant), at the price of an additional computational cost for the median computation stage however.

Rank prediction vs. scoring rule learning When the goal is to rank accurately new unlabeled datasets, rather than to learn a nearly optimal scoring rule explicitly, the following variant of the procedure described above can be considered. Given an unlabeled sample of i.i.d. copies of the input r.v. X $\mathcal{D}_X = \{X_1, \ldots, X_m\}$, instead of aggregating scoring functions s_k defined on the feature space $\mathcal{X}$ and use a consensus rule for ranking the elements of $\mathcal{D}_X$, one may aggregate their restrictions to the finite set $\mathcal{D}_X \subset \mathcal{X}$, or simply the ranks of the unlabeled data as defined by the s_k's.

4 Performance measures for K-partite ranking

We now turn to the main concepts for assessing performance in the K-partite ranking problem. We focus on the notion of ROC surface and Volume Under the ROC Surface (VUS) in the case where $K = 3$ in order to keep the presentation simple. These concepts are generalizations of the well-known ROC curve and AUC criterion which are popular performance measures for bipartite ranking.

4.1 ROC surface

Given a scoring rule $s : \mathbb{R}^d \to \mathbb{R}$, the ROC surface offers a visual display which reflects how the conditional distributions of $s(X)$ given the class label $Y = k$ are separated between each other as $k = 1, \ldots, K$. We introduce the notation $F_{s,k}$ for the cumulative distribution function (cdf) over the real line $\mathbb{R}$ of the random variable $s(X)$ given the class label $Y = k$:

$$\forall t \in \mathbb{R}, \quad F_{s,k}(t) = \mathbb{P}\big\{s(x) \leq t \mid Y = k\big\}.$$

Definition 6 (ROC SURFACE) Let $K \geq 2$. The ROC surface of a real-valued scoring rule s is defined as the plot of the continuous extension of the parametric surface in the unit cube $[0,1]^K$:

$$\begin{aligned} \Delta &\to [0,1]^K \\ (t_1, \ldots, t_{K-1}) &\mapsto \big(F_{s,1}(t_1), F_{s,2}(t_2) - F_{s,2}(t_1), \ldots, 1 - F_{s,K}(t_{K-1})\big), \end{aligned}$$

where $\Delta = \{(t_1, \ldots, t_{K-1}) \in \mathbb{R}^{K-1} : t_1 < \cdots < t_{K-1}\}$.

By "continuous extension", it is meant that discontinuity points, due to jumps or flat parts in the cdfs $F_{s,k}$, are connected by linear segments (parts of hyperplanes). The same convention is considered in the definition of the ROC curve in the bipartite case given in Clémençon and Vayatis (2009b). In the case $K = 3$, on which we restrict our attention from now for simplicity (all results stated in the sequel can be straightforwardly extended to the general situation), the ROC surface thus corresponds to a continuous manifold of dimension 2 in the unit cube of $\mathbb{R}^3$. We also point out that the ROC surface contains the ROC curves of the pairwise problems (f_1, f_2), (f_2, f_3) and (f_1, f_3) which can be obtained as the intersections of the ROC surface with planes orthogonal to each of the axis of the unit cube.

In order to keep track of the relationship between the ROC surface and its sections, we introduce the following notation:

$$\forall \alpha \in [0,1], \quad \mathrm{ROC}_{f_k, f_{k+1}}(s, \alpha) = 1 - F_{s,k+1} \circ F_{s,k}^{-1}(1-\alpha),$$

where we have used the following definition of the generalized inverse of a cdf F: $F^{-1}(u) = \inf\{t \in]-\infty, +\infty] : F(t) \geq u\}$, $u \in [0,1]$.

Proposition 4 (CHANGE OF PARAMETERIZATION) *The ROC surface of the scoring rule s can be obtained as the plot of the continuous extension of the parametric surface*:

$$[0,1]^2 \to \mathbb{R}^3$$
$$(\alpha, \gamma) \mapsto \big(\alpha, \mathrm{ROC}(s, \alpha, \gamma), \gamma\big)$$

where

$$\mathrm{ROC}(s, \alpha, \gamma) = \big(F_{s,2} \circ F_{s,3}^{-1}(1-\gamma) - F_{s,2} \circ F_{s,1}^{-1}(\alpha)\big)_+ \tag{2}$$
$$= \big(\mathrm{ROC}_{f_1, f_2}(s, 1-\alpha) - \mathrm{ROC}_{f_3, f_2}(s, \gamma)\big)_+, \tag{3}$$

with the notation $u_+ = \max(0, u)$, for any real number u.

We point out that, in the case where s has no capacity to discriminate between the three distributions, *i.e.* when $F_{s,1} = F_{s,2} = F_{s,3}$, the ROC surface boils down to the surface delimited by the triangle that connects the points $(1,0,0)$, $(0,1,0)$ and $(0,0,1)$, we then have $\mathrm{ROC}(s, \alpha, \gamma) = 1 - \alpha - \gamma$. By contrast, in the separable situation (see Sect. 2.4), the optimal ROC surface coincides with the surface of the unit cube $[0,1]^3$.

The next lemma characterizes the support of the function whose plot corresponds to the ROC surface (see Fig. 2).

Lemma 1 *For any $(\alpha, \gamma) \in [0,1]^2$, the following statements are equivalent*:

1. $\mathrm{ROC}(s, \alpha, \gamma) > 0$
2. $\mathrm{ROC}_{f_1, f_3}(s, 1-\alpha) > \gamma$.

Other notions of ROC surface have been considered in the literature, depending on the learning problem considered and the goal pursued. In the context of multi-class pattern recognition, they provide a visual display of classification accuracy, as in Ferri et al. (2003) (see also Fieldsend and Everson 2005, 2006 and Hand and Till 2001) from a *one-versus-one* angle or in Flach (2004) when adopting the *one-versus-all* approach. The concept of ROC analysis described above is more adapted to the situation where a natural order on the set of labels exists, just like in ordinal regression, see Waegeman et al. (2008b).

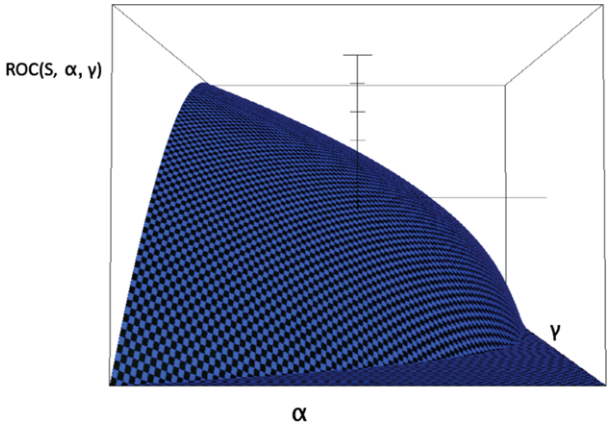


Fig. 2 Plot of the ROC surface of a scoring function

4.2 ROC-optimality and optimal scoring rules

The ROC surface provides a visual tool for assessing ranking performance of a scoring rule. The next theorem provides a formal statement to justify this practice.

Theorem 1 *The following statements are equivalent*:

1. *Assumption 1 is fulfilled and* s^* *is an optimal scoring rule in the sense of Definition 1.*
2. *We have, for any scoring rule* s *and for all* $(\alpha, \gamma) \in [0, 1]^2$,

$$\mathrm{ROC}(s, \alpha, \gamma) \leq \mathrm{ROC}\big(s^*, \alpha, \gamma\big).$$

A nontrivial byproduct of the proof of the previous theorem is that optimizing the ROC surface amounts to simultaneously optimizing the ROC curves related to the two pairs of distributions (f_1, f_2) and (f_2, f_3).

The theorem indicates that optimality for scoring rules in the sense of Definition 1 is equivalent to optimality in the sense of the ROC surface. Therefore, the ROC surface provides a complete characterization of the ranking performance of a scoring rule in the K-partite problem.

We now introduce the following notations: for any $\alpha \in [0, 1]$ and any scoring rule s,

- the quantile of order $(1 - \alpha)$ of the conditional distribution of the random variable $s(X)$ given $Y = k$:

$$Q^{(k)}(s, \alpha) = F_{s,k}^{-1}(1 - \alpha),$$

- the level set of the scoring rule s with the top elements of class $Y = k$:

$$R_{s,\alpha}^{(k)} = \big\{x \in \mathcal{X} | s(x) > Q^{(k)}(s, \alpha)\big\}.$$

Proposition 5 *Suppose that Assumption 1 is fulfilled and consider* s^* *an optimal scoring rule in the sense of Definition 1. Also assume that* $\eta(X)$ *is a continuous random variable, then we have*: $\forall(\alpha, \gamma) \in [0, 1]^2$:

$$\mathrm{ROC}^*\big(s^*, \alpha, \gamma\big) - \mathrm{ROC}(s, \alpha, \gamma) \leq \mathbb{I}\big\{\gamma \leq \mathrm{ROC}_{F_1,F_3}\big(s^*, 1 - \alpha\big)\big\} \cdot \big(\Theta_1(s, \alpha) + \Theta_2(s, \gamma)\big),$$

where

$$\Theta_1(s,\alpha) = \frac{\mathbb{I}\{\alpha \neq 0\}}{p_2 Q^{(1)}(\eta_1,\alpha)} \mathbb{E}\big[\big|\eta_1(x) - Q^{(1)}(\eta_1,\alpha)\big| \cdot \mathbb{I}\big\{R^{(1)}_{s^*,\alpha} \Delta R^{(1)}_{s,\alpha}\big\}\big],$$

$$\Theta_2(s,\gamma) = \frac{\mathbb{I}\{\gamma \neq 1\}}{p_2 Q^{(3)}(\eta_3,1-\gamma)} \mathbb{E}\big[\big|\eta_3(X) - Q^{(3)}(\eta_3,1-\gamma)\big| \cdot \mathbb{I}\big\{R^{(3)}_{s^*,1-\gamma} \Delta R^{(3)}_{s,1-\gamma}\big\}\big].$$

We have used the notation $A \Delta B = (A \setminus B) \cup (B \setminus A)$ for the symmetric difference between sets A and B.

The previous proposition provides a key inequality for the statistical results developed in the sequel.

4.3 Volume Under the ROC Surface (VUS)

In the bipartite case, a standard summary of ranking performance is the Area Under an ROC Curve (or AUC). In a similar manner, one may consider the *volume under the* ROC *surface* (VUS in abbreviated form) in the three-class framework. We follow here Scurfield (1996) but we mention that other notions of ROC surface can be found in the literature, leading to other summary quantities, also referred to as VUS, such as those introduced in Hand and Till (2001).

Definition 7 (Volume Under the ROC Surface) We define the VUS of a real-valued scoring rule s as:

$$\mathrm{VUS}(s) = \int_0^1 \int_0^1 \mathrm{ROC}(s,\alpha,\gamma)\,\mathrm{d}\alpha\,\mathrm{d}\gamma.$$

An alternative expression of VUS can be derived with a change of parameters:

$$\mathrm{VUS}(s) = \int_0^1 \mathrm{ROC}_{f_1,f_2}(s,1-\alpha)\mathrm{ROC}_{f_1,f_3}(s,1-\alpha)\,\mathrm{d}\alpha - \int_0^1 \mathrm{ROC}_{f_3,f_2}(s,\gamma)\big(1-\mathrm{ROC}_{f_3,f_1}(s,\gamma)\big)\,\mathrm{d}\gamma.$$

The next proposition describes two extreme cases.

Proposition 6 *Consider a real-valued scoring rule s.*

1. *If $F_{s,1} = F_{s,2} = F_{s,3}$, then* $\mathrm{VUS}(s) = 1/6$.
2. *If the density functions of $F_{s,1}$, $F_{s,2}$, $F_{s,3}$ have disjoint supports, then* $\mathrm{VUS}(s) = 1$.

Like the AUC criterion, the VUS can be interpreted in a probabilistic manner. For completeness, we recall the following result.

Proposition 7 (Scurfield 1996) *For any scoring function $s \in \mathcal{S}$, we have*:

$$\mathrm{VUS}(s) = \mathbb{P}\big\{s(X_1) < s(X_2) < s(X_3) | Y_1 = 1, Y_2 = 2, Y_3 = 3\big\} + \frac{1}{2}\mathbb{P}\big\{s(X_1) = s(X_2) < s(X_3) | Y_1 = 1, Y_2 = 2, Y_3 = 3\big\}$$

$$+\frac{1}{2}\mathbb{P}\big\{s(X_1)<s(X_2)=s(X_3)|Y_1=1,Y_2=2,Y_3=3\big\}$$
$$+\frac{1}{6}\mathbb{P}\big\{s(X_1)=s(X_2)=s(X_3)|Y_1=1,Y_2=2,Y_3=3\big\},$$

where (X_1,Y_1), (X_2,Y_2) *and* (X_3,Y_3) *denote independent copies of the random pair* (X,Y).

In the case where the distribution of s is continuous, the last three terms in the term on the right hand side vanish and the VUS boils down to the probability that, given three random instances X_1, X_2 and X_3 with respective labels $Y_1=1$, $Y_2=2$ and $Y_3=3$, the scoring rule s ranks them in the right order.

4.4 VUS-optimality

We now consider the notion of optimality with respect to the VUS criterion and provide expressions of the deficit of VUS for any scoring rule which highlight the connection with AUC maximizers for the bipartite subproblems.

Proposition 8 (VUS OPTIMALITY) *Under Assumption 1, we have, for any real-valued scoring rule* s *and any optimal scoring rule* s^*:

$$\mathrm{VUS}(s)\leq \mathrm{VUS}\big(s^*\big).$$

We denote the maximal value of the VUS by $\mathrm{VUS}^*=\mathrm{VUS}(s^*)$

This result shows that optimal scoring rules in the sense of Definition 1 coincide with optimal elements in the sense of VUS. This simple statement grounds the use of empirical VUC maximization strategies for the K-partite ranking problem.

When the Assumption 1 is not fulfilled, the VUS can still be used as a performance criterion, both in the multiclass classification context (Landgrebe and Duin 2006; Ferri et al. 2003) and in the ordinal regression setup (Waegeman et al. 2008b). However, the interpretation of maximizers of VUS as optimal orderings is highly questionable. For instance, in the situation described in Example 1, one may easily check that, when $\omega_{1,1}=4/11$, $\omega_{1,2}=6/11$, $\omega_{1,3}=\omega_{3,1}=1/11$, $\omega_{2,1}=\omega_{2,2}=3/11$ and $\omega_{2,3}=\omega_{3,2}=\omega_{3,3}=5/11$, the maximum VUS (equal to 0.2543) is reached by the scoring rule corresponding to strict orders $\prec$ and $\prec'$, such that $x_3\prec x_2\prec x_1$ and $x_2\prec' x_3\prec' x_1$ respectively, both at the same time.

We introduce the definition for the AUC of the bipartite ranking problem with the pair of distributions (f_k,f_{k+1}):

Definition 8 (AUC) Let X_1 and X_2 independent random variables with distribution f_k and f_{k+1} respectively. We set:

$$\mathrm{AUC}_{f_k,f_{k+1}}(s)=\mathbb{P}\big\{s(X_1)<s(X_2)\big\}+\frac{1}{2}\mathbb{P}\big\{s(X_1)=s(X_2)\big\}.$$

We now state the result which establishes the relevance of AUC as an optimality criterion for the bipartite ranking problem.

Proposition 9 *Fix $k \in \{1, \ldots, K-1\}$ and consider s_k^* a pairwise optimal scoring rule according to Definition 4. Then we have, for any scoring rule*:

$$\mathrm{AUC}_{f_k, f_{k+1}}(s) \leq \mathrm{AUC}_{f_k, f_{k+1}}\big(s_k^*\big).$$

Moreover, we denote the maximal value of the AUC for the bipartite (f_k, f_{k+1}) ranking problem by: $\mathrm{AUC}^*_{f_k, f_{k+1}} = \mathrm{AUC}_{f_k, f_{k+1}}(s_k^*)$.

The next result makes clear that if a scoring rule s solves simultaneously all the bipartite ranking subproblems then it also solves the global K-partite ranking problem. For simplicity, we present the result in the case $K = 3$.

Theorem 2 (DEFICIT OF VUS) *Suppose that Assumption 1 is fulfilled. Then, for any scoring rule s and any optimal scoring rule s^*, we have*

$$\begin{aligned}\mathrm{VUS}\big(s^*\big) - \mathrm{VUS}(s) &\leq \big(\mathrm{AUC}^*_{f_1,f_2} - \mathrm{AUC}_{f_1,f_2}(s)\big) + \big(\mathrm{AUC}^*_{f_2,f_3} - \mathrm{AUC}_{f_2,f_3}(s)\big) \\ &\leq \frac{2}{3}\Big(\big(\mathrm{AUC}^*_{f_1,f_2} - \mathrm{AUC}_{f_1,f_2}(s)\big) \\ &\quad + \big(\mathrm{AUC}^*_{f_2,f3} - \mathrm{AUC}_{f_2,f_3}(s)\big) + \big(\mathrm{AUC}^*_{f_1,f_3} - \mathrm{AUC}_{f_1,f_3}(s)\big)\Big).\end{aligned}$$

5 Consistency of pairwise aggregation and other strategies for K-partite ranking

5.1 Definition of VUS-consistency and main result

In this section, we assume a data sample $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ is available and composed by n i.i.d. copies of the random pair (X, Y). Our goal here is to learn from the sample $\mathcal{D}_n$ how to build a real-valued scoring rule $\widehat{s}_n$ such that its ROC surface is as close as possible to the optimal ROC surface. We propose to consider a weak concept of consistency which relies on the VUS.

Definition 9 (VUS-CONSISTENCY) Suppose that Assumption 1 is fulfilled. Let $(s_n)_{n\geq 1}$ be a sequence of random scoring rules on $\mathbb{R}^d$, then:

– the sequence $\{s_n\}$ is called VUS-consistent if

$$\mathrm{VUS}^* - \mathrm{VUS}(s_n) \to 0 \quad \text{in probability,}$$

– the sequence $\{s_n\}$ is called strongly VUS-consistent if

$$\mathrm{VUS}^* - \mathrm{VUS}(s_n) \to 0 \quad \text{with probability one.}$$

Remark 1 We note that the deficit of VUS can be interpreted as an L_1 distance between ROC surfaces of s_n and s^*:

$$\mathrm{VUS}^* - \mathrm{VUS}(s_n) = \int\int_{(\alpha,\gamma)\in[0,1]^2} \big|\mathrm{ROC}\big(s^*, \alpha, \gamma\big) - \mathrm{ROC}(s_n, \alpha, \gamma)\big|\, \mathrm{d}\alpha\, \mathrm{d}\gamma,$$

and in this sense the notion of consistency is weak. Indeed, a stronger sense of consistency could be given by considering the supremum norm between surfaces:

$$d_\infty\big(s^*, s_n\big) = \sup_{(\alpha,\gamma)\in[0,1]^2} \big|\mathrm{ROC}\big(s^*, \alpha, \gamma\big) - \mathrm{ROC}(s_n, \alpha, \gamma)\big|.$$

The study of accuracy of K-partite ranking methods in this sense is beyond the scope of the present paper (in contrast to the L_1 norm, the quantity $d_\infty(s^*, s)$ cannot be decomposed in an additive manner). Extensions of bipartite ranking procedures such as the TREERANK and the RANKOVER algorithms (see Clémençon and Vayatis 2009b and 2010), for which consistency in supremum norm is guaranteed under some specific assumptions, will be considered in future work.

In order to state the main result, we need an additional assumption on the distribution of the random pair (X, Y). The reason why this assumption is needed will be explained in the next section.

Assumption 2 For all $k \in \{1, \ldots, K-1\}$, the (pairwise) posterior probability given by $\eta_{k+1}(X)/(\eta_k(X) + \eta_{k+1}(X))$ is a continuous random variable and there exist $c < \infty$ and $a \in (0, 1)$ such that

$$\forall x \in \mathcal{X}, \quad \mathbb{E}\left[\left|\frac{\eta_{k+1}(X)}{\eta_{k+1}(X)+\eta_k(X)} - \frac{\eta_{k+1}(x)}{\eta_{k+1}(x)+\eta_k(x)}\right|^{-a}\right] \leq c. \tag{4}$$

In the statistical learning literature, Assumption 2 is referred to as the noise condition and goes back to the work of Tsybakov (2004). It has been adapted to the framework of bipartite ranking in Clémençon et al. (2008). For completeness, we state a result from this latter paper (see Corollary 8 within) which offers a simple sufficient condition for the Assumption 2 to be fulfilled.

Proposition 10 *If the distribution of the r.v.* $\eta_{k+1}(X)/(\eta_k(X) + \eta_{k+1}(X))$ *has a bounded density, then Assumption 2 is satisfied.*

We will also need to use the notion of AUC consistency for the bipartite ranking subproblems.

Definition 10 (AUC CONSISTENCY) For k fixed in $\{1, \ldots, K-1\}$, a sequence $(s_n)_{n\geq 1}$ of scoring rules is said to be AUC-consistent (respectively, strongly AUC-consistent) for the bipartite problem (f_k, f_{k+1}) if it satisfies:

$$\mathrm{AUC}_{f_k, f_{k+1}}(s_n) \to \mathrm{AUC}^*_{f_k, f_{k+1}} \text{ in probability (resp., with probability one).}$$

We can now state the main consistency result of the paper which concerns the Kendall aggregation procedure described in Sect. 3.2. Indeed, the following theorem reveals that the notion of median scoring rule introduced in Definition 3 preserves AUC consistency for bipartite subproblems and thus yields a VUS consistent scoring rule for the K-partite problem. It is assumed that the solutions to the bipartite subproblems are AUC-consistent for each specific pair of class distributions (f_k, f_{k+1}), $1 \leq k < K$. For simplicity, we formulate the result in the case $K = 3$.

Theorem 3 *We consider a class of candidate scoring rules* $\mathcal{S}_1$, $(s_n^{(1)})_{n\geq 1}$, $(s_n^{(2)})_{n\geq 1}$ *two sequences of scoring rules in* $\mathcal{S}_1$. *We use the notation* $\Sigma_{2,n} = \{s_n^{(1)}, s_n^{(2)}\}$. *Assume the following*:

1. *Assumptions* 1 *and* 2 *hold true*.
2. *The class* $\mathcal{S}_1$ *contains an optimal scoring rule*.
3. *The sequences* $(s_n^{(1)})_{n\geq 1}$ *and* $(s_n^{(2)})_{n\geq 1}$ *are* (*strongly*) AUC-*consistent for the bipartite ranking subproblems related to the pairs of distributions* (f_1, f_2) *and* (f_2, f_3) *respectively*.
4. *Assume that*, *for all* n, *there exists a median scoring rule* $\overline{s}_n$ *in the sense of Definition* 3 *with respect to* $(\mathcal{S}_1, \Sigma_{2,n})$.

Then the median scoring rule $\overline{s}_n$ *is* (*strongly*) VUS-*consistent*.

Discussion The first assumption of Theorem 3 puts a restriction on the class of distributions for which such a consistency result holds. Assumption 1 actually guarantees that the very problem of K-partite makes sense and the existence of an optimal scoring rule. Assumption 2 can be seen as a "light" restriction since it still covers a large class of distributions commonly used in probabilistic modeling. The third and fourth assumptions are natural as we expect first to have efficient solutions to the bipartite subproblems before considering reasonable solutions to the K-partite problem. The most restrictive assumption is definitely the second one about the fact that the class of candidates contains an optimal element. Indeed, it is easy to weaken this assumption at the price of an additional bias term by assuming that the scoring rules $s_n^{(1)}$, $s_n^{(2)}$ and $\overline{s}_n$ belong to a set $\mathcal{S}_1^{(n)}$, such that there exists a sequence $(s_n^*)_{n\geq 1}$ with $s_n^* \in \mathcal{S}_1^{(n)}$ and $\mathrm{VUS}(s_n^*) \to \mathrm{VUS}^*$ as $n \to \infty$. We decided not to include this refinement as this is merely a technical argument which does not offer additional insights on the nature of the problem.

5.2 From AUC consistency to VUS consistency

In this section, we introduce auxiliary results which contribute to the proof of the main theorem (details are provided in the Appendix). Key arguments rely on the relationship between the solutions of the bipartite ranking subproblems and those of the K-partite problem. In particular, a sequence of scoring rules that is simultaneously AUC-consistent for the bipartite ranking problems related to the two pairs of distributions (f_1, f_2) and (f_2, f_3) is VUS-consistent. Indeed, we have the following corollary.

Corollary 1 *Suppose that Assumption* 1 *is satisfied*. *Let* $(s_n)_{n\geq 1}$ *be a sequence of scoring rules*. *The following assertions are equivalent*.

(i) *The sequence* $(s_n)_n$ *of scoring rules is* (*strongly*) *VUS-optimal*.
(ii) *We have simultaneously when* $n \to \infty$:

$$\mathrm{AUC}_{f_1,f_2}(s_n) \to \mathrm{AUC}^*_{f_1,f_2}$$
$$\mathrm{AUC}_{f_2,f_3}(s_n) \to \mathrm{AUC}^*_{f_2,f_3}$$

(*with probability one*) *in probability*.

It follows from this result that the 3-partite ranking problem can be cast in terms of a double-criterion optimization task, consisting in finding a scoring rule s that simultaneously

maximizes $\mathrm{AUC}_{f_1,f_2}(s)$ and $\mathrm{AUC}_{f_2,f_3}(s)$. This result provides a theoretical basis for the justification of our pairwise aggregation procedure. We mention that the idea of decomposing the K-partite ranking into several bipartite ranking subproblems has also been considered in Fürnkranz et al. (2009) but the aggregation stage is performed with a different strategy.

The other type of result which is needed concerns the connection between the aggregation principle based on a consensus approach (Kendall τ) and the performance metrics involved in the K-partite ranking problem. The next results establish inequalities which relate the AUC and the Kendall τ in a quantitative manner.

Proposition 11 *Let p be a real number in $(0,1)$. Consider two probability distributions f_k and f_k+1 on the set $\mathcal{X}$. We assume that the distribution of X comes from the mixture with density function given by $(1-p)f_k + pf_{k+1}$. For any real-valued scoring rules s_1 and s_2 on $\mathbb{R}^d$, we have*:

$$\left|\mathrm{AUC}_{f_k,f_{k+1}}(s_1) - \mathrm{AUC}_{f_k,f_{k+1}}(s_2)\right| \leq \frac{1-\tau(s_1,s_2)}{4p(1-p)}.$$

We point out that it is generally vain to look for a reverse control: indeed, scoring functions yielding different rankings may have exactly the same AUC. However, the following result guarantees that a scoring function with a nearly optimal AUC is close to optimal scoring functions in a certain sense, under the additional assumption that the noise condition introduced in Clémençon et al. (2008) is fulfilled.

Proposition 12 *Under Assumption 2, we have, for any $k \in \{1,\ldots,K-1\}$, for any scoring rule s and any pairwise optimal scoring rule s_k^**:

$$1-\tau\left(s_k^*, s\right) \leq C \cdot \left(\mathrm{AUC}^*_{f_k,f_{k+1}} - \mathrm{AUC}_{f_k,f_{k+1}}(s)\right)^{a/(1+a)},$$

with $C = 3c^{1/(1+a)} \cdot (2p_k p_{k+1})^{a/(1+a)}$.

5.3 Alternative approaches to K-partite ranking

In this section, we also mention, for completeness, two other approaches to K-partite ranking.

5.3.1 Empirical VUS *maximization*

The first approach extends the popular principle of empirical risk minimization, see Vapnik (1999). For K-partite ranking, this programme has been carried out in Rajaram and Agarwal (2005) with an accuracy measure based on the loss function $(Y-Y')_+^{\xi}(\mathbb{I}\{s(X)<s(X')\} + (1/2)\cdot\mathbb{I}\{s(X)=s(X')\})$, with $\xi \geq 0$. In our setup, the idea would be to optimize a statistical counterpart of the unknown functional VUS(.) over a class $\mathcal{S}_1$ of candidate scoring rules. Based on the training dataset $\mathcal{D}_n$, a natural empirical counterpart of $\mathrm{VUS}(s)$ is the three-sample U-statistic

$$\widehat{\mathrm{VUS}}_n(s) = \frac{1}{n_1 n_2 n_3} \sum_{1\leq i,j,k\leq n} h_s(X_i, X_j, X_k)\cdot \mathbb{I}\{Y_i=1, Y_j=2, Y_k=3\}, \tag{5}$$

with kernel given by

$$h_s(x_1,x_2,x_3) = \mathbb{I}\big\{s(x_1) < s(x_2) < s(x_3)\big\} + \frac{1}{2}\mathbb{I}\big\{s(x_1) = s(x_2) < s(x_3)\big\} + \frac{1}{2}\mathbb{I}\big\{s(x_1) < s(x_2) = s(x_3)\big\} + \frac{1}{6}\mathbb{I}\big\{s(x_1) = s(x_2) = s(x_3)\big\},$$

for any $(x_1, x_2, x_3) \in \mathcal{X}^3$. The computational complexity of empirical VUS calculation is investigated in Waegeman et al. (2008a).

The theoretical analysis shall rely on concentration properties of U-processes in order to control the deviation between the empirical and theoretical versions of the VUS criterion uniformly over the class $\mathcal{S}_1$. Such an analysis was performed in the bipartite case in Clémençon et al. (2008) and we expect that it can be extended in the K-partite case. In contrast, algorithmic aspects of the issue of maximizing the empirical VUS criterion (or a concave surrogate) are much less straightforward and the question of extending optimization strategies such as those introduced in Clémençon and Vayatis (2009b) or Clémençon and Vayatis (2010) requires, for instance, significant methodological progress.

5.3.2 Plug-in scoring rule

As shown by Proposition 1, when Assumption 1 is fulfilled, the regression function η is an optimal scoring function. The *plug-in* approach consists of estimating the latter and use the resulting estimate as a scoring rule. For instance, one may estimate the posterior probabilities $(\eta_1(x), \ldots, \eta_K(x))$ by an empirical counterpart $(\widehat{\eta}_1(x), \ldots, \widehat{\eta}_K(x))$ based on the training data and consider the ordering on $\mathbb{R}^d$ induced by the estimator $\widehat{\eta}(x) = \sum_{k=1}^K k\widehat{\eta}_k(x)$. We refer to Clémençon and Vayatis (2009a) and Clémençon and Robbiano (2011) for preliminary theoretical results based on this strategy in the bipartite context and Audibert and Tsybakov (2007) for an account of the plug-in approach in binary classification. It is expected that an accurate estimate of $\eta(x)$ will define a ranking rule similar to the optimal one, with nearly maximal VUS. As an illustration of this approach, the next result relates the *deficit of* VUS of a scoring function $\widehat{\eta}$ to its $L_1(\mu)$-error as an estimate of η. We assume for simplicity that all class-conditional distributions have the same support.

Proposition 13 *Suppose that Assumption 1 is fulfilled. Let $\widehat{\eta}$ be an approximant of η. Assume that both the random variables $\eta(X)$ and $\widehat{\eta}(X)$ are continuous. We have*:

$$\mathrm{VUS}^* - \mathrm{VUS}(\widehat{\eta}) \leq \frac{p_1 + p_3}{p_1 p_2 p_3} \cdot \mathbb{E}\big[\big|\eta(X) - \widehat{\eta}(X)\big|\big].$$

This result reveals that a $L_1(\mu)$-consistent estimator, *i.e.* an estimator $\widehat{\eta}_n$ such that $\mathbb{E}[|\eta(X) - \widehat{\eta}_n(X)|]$ converges to zero in probability as $n \to \infty$, yields a VUS-consistent ranking procedure. However, from a practical perspective, such procedures should be avoided when dealing with high-dimensional data, since they are obviously confronted to the curse of dimensionality.

5.4 Connections with regression estimation and ordinal regression

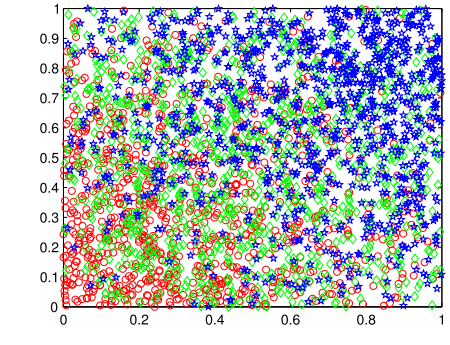
Whereas standard multi-class classification ignores the possible ordinal structure of the output space, ordinal regression takes the latter into account by penalizing more and more the error of a classifier candidate C on an example (X, Y) as $|C(X) - Y|$ increases. In

general, the loss function chosen is of the form $\psi(c, y) = \Psi(|c - y|)$, $(c, y) \in \{1, \ldots, K\}^2$, where $\Psi : \{0, \ldots, K-1\} \to \mathbb{R}_+$ is some nondecreasing mapping. The most commonly used choice is $\Psi(u) = u$, corresponding to the risk $L(C) = \mathbb{E}[|C(X) - Y|]$, referred to as the *expected ordinal regression error* sometimes, *cf.* Agarwal (2008). In this case, it is shown that the optimal classifier can be built by thresholding the regression function at specific levels $t_0 = 0 < t_1^* < \cdots < t_{K-1}^* < 1 = t_K$, that it so say it is of the form $C^*(x) = \sum_{k=1}^K k \cdot \mathbb{I}\{t_{k-1}^* \leq \eta(x) < t_k^*\}$ when assuming that $\eta(X) = \mathbb{E}[Y \mid X]$ is a continuous r.v. for simplicity. Based on this observation, a popular approach to ordinal regression lies in estimating first the regression function η by an empirical counterpart $\widehat{\eta}$ (through minimization of an estimate of $R(f) = \mathbb{E}[(Y - f(X))^2]$ over a specific class $\mathcal{F}$ of function candidates f, in general) and choosing next a collection $\mathbf{t}$ of thresholds $t_0 = 0 < t_1 < \cdots < t_{K-1} < 1 = t_K$ in order to minimize a statistical version of $L(C_{\mathbf{t}})$ where $C_{\mathbf{t}}(x) = \sum_{k=1}^K k \cdot \mathbb{I}\{t_{k-1} \leq \widehat{\eta}(x) < t_k\}$. Such procedures are sometimes termed *regression-based algorithms*, see Agarwal (2008). One may refer to Kramer et al. (2001) in the case of regression trees for instance.

6 Illustrative numerical experiments

It is the purpose of this section to illustrate the approach described above by numerical results and provide some empirical evidence for its efficacy. Since our goal is here to show that, beyond its theoretical validity, the Kendall aggregation approach to multi-class ranking actually works in practice, rather than to provide a detailed empirical study of its performance on benchmark artificial/real datasets compared to that of possible competitors (this will be the subject of a forthcoming paper), in the subsequent experimental analysis we have considered two simple data generative models, for which one may easily check Assumption 1 and compute the optimal ROC surface (as well as the optimum value VUS*), which the results obtained must be compared to. The first example involves mixtures of Gaussian distributions, while the second one is based on mixtures of uniform distributions, the target ROC surface being piecewise linear in the latter case (*cf.* assertion 4 in Proposition 14). Here, the artificial data simulated are split into a *training sample* and a *test sample*, used for plotting the "test ROC surfaces".

The learning algorithm used for solving the bipartite ranking subproblems at the first stage of the procedure is the TreeRank procedure based on locally weighted versions of the CART method (with axis parallel splits), see Clémençon et al. (2011a) for a detailed description of the algorithm (as well as Clémençon and Vayatis 2009b for rigorous statistical foundations of this method). Precisely, we used a package for R statistical software (see http://www.r-project.org) implementing TreeRank (with the "default" parameters: minsplit = (size of training sample)/20, maxdepth = 10, mincrit = 0), available at http://treerank.sourceforge.net, see Baskiotis et al. (2010). The scoring rules produced at stage 1 are thus (tree-structured and) piecewise constant, making the aggregating procedure described in Sect. 3.2 quite feasible. Indeed, if $s_1, \ldots, s_M$ are scoring functions that are all constant on the cells of a finite partition $\mathcal{P}$ of the input space $\mathcal{X}$, one easily see that the infimum $\inf_{s \in \mathcal{S}_0} \sum_{m=1}^M d_{\tau_\mu}(s, s_m)$ reduces to a minimum over a finite collection of scoring functions that are also constant on $\mathcal{P}$'s cells and is thus attained. As underlined in Sect. 3.2, when the number of cells is large, median computation may become practically unfeasible and the use of a meta-heuristic can be then considered for approximation purpose (simulated annealing, tabu search, *etc.*), here the ranking obtained by taking the mean ranks over the $K - 1$ rankings of the test data has been improved in the Kendall consensus sense by means of a standard simulated annealing technique.



a. Pooled sample: red circles represent instances with label "1", green diamonds those with label "2" and blue stars those with label "3"

b. Optimal level sets.

Fig. 3 First example—mixture of Gaussian distributions

For comparison purpose, we have also implemented two ranking algorithms, RankBoost (when aggregating 30 stumps, see Rudin et al. 2005) and SVMRank (with linear and Gaussian kernels with respective parameters $C = 20$ and $(C, \gamma) = (0.01)$, see Herbrich et al. 2000), using the SVM-light implementation available at http://svmlight.joachims.org/. We have also used the RankRLS method (http://www.tucs.fi/RLScore, see Pahikkala et al. 2007) that implements a regularized least square algorithm with linear kernel ("*bias* = 1") and with Gaussian kernel ($\gamma = 0.01$), selection of the intercept on a grid being performed through a leave-one-out procedure. For completeness, the Kendall aggregation procedure has also been implemented with RankBoost for solving the bipartite subproblems.

First example (mixtures of Gaussian distributions) Consider a q-dimensional Gaussian random vector Z, drawn as $\mathcal{N}(\mu, \Gamma)$, and a Borelian set $C \subset \mathbb{R}^q$ weighted by $\mathcal{N}(\mu, \Gamma)$. We denote by $\mathcal{N}_C(\mu, \Gamma)$ the conditional distribution of Z given $Z \in C$. Equipped with this notation, we can write the class distributions used in this example as:

$$f_1(x) = \mathcal{N}_{[0,1]^2}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1/4 & 0\\0 & 1/4\end{pmatrix}\right)$$

$$f_2(x) = \mathcal{N}_{[0,1]^2}\left(\begin{pmatrix}1/2\\1/2\end{pmatrix}, \begin{pmatrix}1/4 & 0\\0 & 1/4\end{pmatrix}\right)$$

$$f_3(x) = \mathcal{N}_{[0,1]^2}\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1/4 & 0\\0 & 1/4\end{pmatrix}\right).$$

When $p_1 = p_2 = p_3 = 1/3$, the regression function is then an increasing transform of $(x_1, x_2) \in [0, 1]^2 \mapsto x_1 + x_2$, it is given by:

$$\eta(x) = \frac{2.79 \cdot e^{-(x_1+x_2)^2} + 2 \cdot 1.37 \cdot e^{-(x_1+x_2-1)^2} + 3 \cdot 2.79 \cdot e^{-(x_1+x_2-2)^2}}{2.79 \cdot e^{-(x_1+x_2)^2} + 1.37 \cdot \exp^{-(x_1+x_2-1)^2} + 2.79 \cdot e^{-(x_1+x_2-2)^2}}.$$

The simulated dataset is plotted in Fig. 3a, while some level sets of the regression function are represented in Fig. 3b. We have drawn 50 training samples of size $n = 3000$ and a test sample of size 3000. Using TREERANK, we learn 3 bipartite ranking rules: $s^{(1)}(x)$ based on data with labels "1" and "2", $s^{(2)}(x)$ based on data with labels "2" and "3" and

Table 1 Comparison of the VUS: "Gaussian" experiment—$\mathrm{VUS}^* = 0.4369$

Method	$\overline{\mathrm{VUS}}(\widehat{\sigma})$
TreeRank 1v2	0.3703 ($\pm$0.0102)
TreeRank 2v3	0.3728 ($\pm$0.0104)
TreeRank 1v3	0.3972 ($\pm$0.0053)
TreeRank Agg	0.4118 ($\pm$0.0054)
RankBoostVUS	0.4281 ($\pm$0.0024)
RankBoost Agg	0.4305 ($\pm$0.0019)
SVMrank lin	0.4367 ($\pm$0.0003)
SVMrank gauss	0.4363 ($\pm$0.0009)
RLScore lin	0.4368 ($\pm$0.0003)
RLScore gauss	0.4366 ($\pm$0.0006)

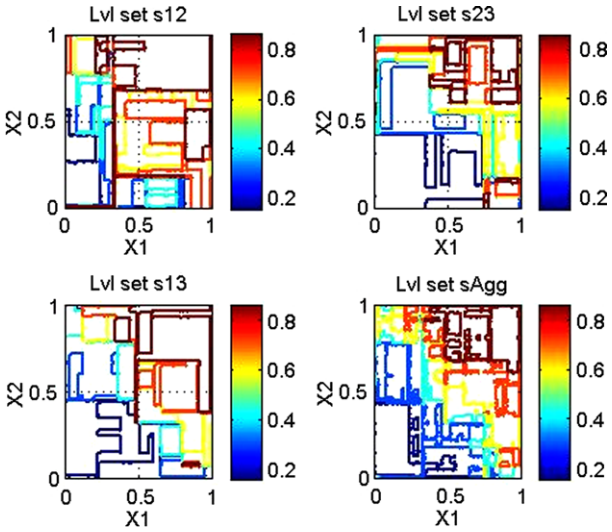


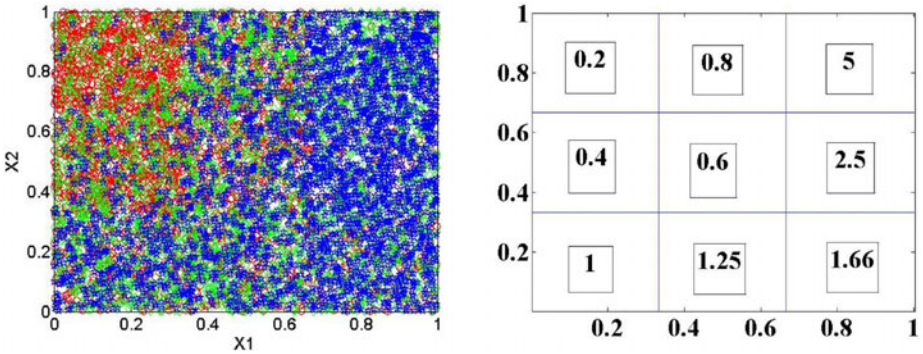
Fig. 4 Levels sets of the scoring functions "TreeRank 1v2", "TreeRank 2v3", "TreeRank 1v3" and "TreeRank Agg" in a top-down left-right manner

$s^{(3)}(x)$ based on data with labels "1" and "3". Finally, $s^{(1)}$ and $s^{(2)}$ are aggregated through the procedure described in Sect. 3.2, yielding the score called "TreeRank Agg" in Table 1. We also used each scoring function separately to rank the test data and compute a test estimate of the VUS ("TreeRank 1v2", "2v3", "1v3"). The scoring function produced by RankBoost is referred to as "RankBoostVUS", while that obtained by Kendall aggregation based on (a bipartite implementation of) RankBoost is called "RankBoost Agg". The scoring rule computed through SVMrank (respectively, through RankRLS) based on a linear and a Gaussian kernels are respectively called "SVMrank lin" and "SVMrank gauss" (respectively, "RLScore lin" and "RLScore gauss"). Averages ($\overline{\mathrm{VUS}}$) over the 50 training samples have been next computed, as well as standard deviations $\widehat{\sigma}$, they are given in Table 1 with the results of the earlier described algorithms. For comparison purpose, some level sets of the TreeRank scoring functions learnt from the first training sample are displayed in Fig. 4.

Second example (mixtures of uniform distributions) The artificial data sample used in this second example is represented in Fig. 5a and has been generated as follows. The unit square $\mathcal{X} = [0, 1]^2$ is split into 9 squares of equal size and we defined next the scoring function s^*

Table 2 Values of the η_k's on each of the nine subsquare of $[0, 1]^2$, *cf.* Fig. 5b

s^*	$s^*_{1,2}$	$s^*_{2,3}$	η_1	η_2	η_3
0.2	0.2	0.2	0.7692	0.2000	0.0308
0.4	0.4	0.2	0.6250	0.3250	0.0500
0.6	0.8	0.6	0.3968	0.4127	0.1905
0.8	0.8	0.8	0.3731	0.3881	0.2388
1	1	1	0.3030	0.3939	0.3030
1.25	1.25	1	0.2581	0.4194	0.3226
1.66	1.66	1.66	0.1682	0.3645	0.4673
2.5	2.5	2.5	0.0952	0.3095	0.5952
5	2.5	5	0.0597	0.1940	0.7463



a. Pooled sample: red circles represent instances with label "1", green diamonds those with label "2" and blue stars those with label "3"

b. Optimal scoring function s^*.

Fig. 5 Second example—mixtures of uniform distributions

as the function constant on each of these squares depicted by Fig. 5b). We then chose the uniform distribution over the unit square as marginal distribution of X and took $\phi_{1,2}(x) = s^*_{1,2}(x)/1.3$ and $\phi_{2,3} = 1.3 \times s^*_{2,3}(x)$. As $s^*_{1,2}$ and $s^*_{2,3}$ are non-decreasing functions of s^* (see Table 2): $\phi_{2,1}$ and $\phi_{3,2}$ are thus non-decreasing functions of s^*, by virtue of Theorem 1, the class distributions check the monotonicity assumption 1. Computation of the η_i's on each part of $\mathcal{X}$ is then straightforward, see Table 2.

Here 50 training samples of size $n = 9000$ plus a test sample of size 9000 have been generated. The performance results are reported in Table 3. For comparison purpose, some level sets of the scoring functions learned on the first training sample for each method are represented in Fig. 6.

Cardiotocography data We also illustrate the methodology promoted in this paper by implementing it on a real data set, the *Cardiotocography Data Set* considered in Frank and Asuncion (2010) namely. The data have been collected as follows: 2126 fetal cardiotocograms (CTG's in abbreviated form) have been automatically processed and the respective diagnostic features measured. The CTG's have been next analyzed by three expert obstetricians and a consensus ordinal label has been then assigned to each of them,

Table 3 Comparison of the VUS: "uniform" experiment—VUS* = 0.3855

Method	$\overline{\text{VUS}}(\widehat{\sigma})$
TreeRank 1v2	0.3681 (±0.0060)
TreeRank 2v3	0.3611 (±0.0056)
TreeRank 1v3	0.3774 (±0.0037)
TreeRank Agg	0.3818 (±0.0027)
RankBoostVUS	0.3681 (±0.0013)
RankBoost Agg	0.3687 (±0.0013)
SVMrank lin	0.3557 (±0.0008)
SVMrank gauss	0.3734 (±0.0008)
RLScore lin	0.3554 (±0.0005)
RLScore gauss	0.3742 (±0.0007)

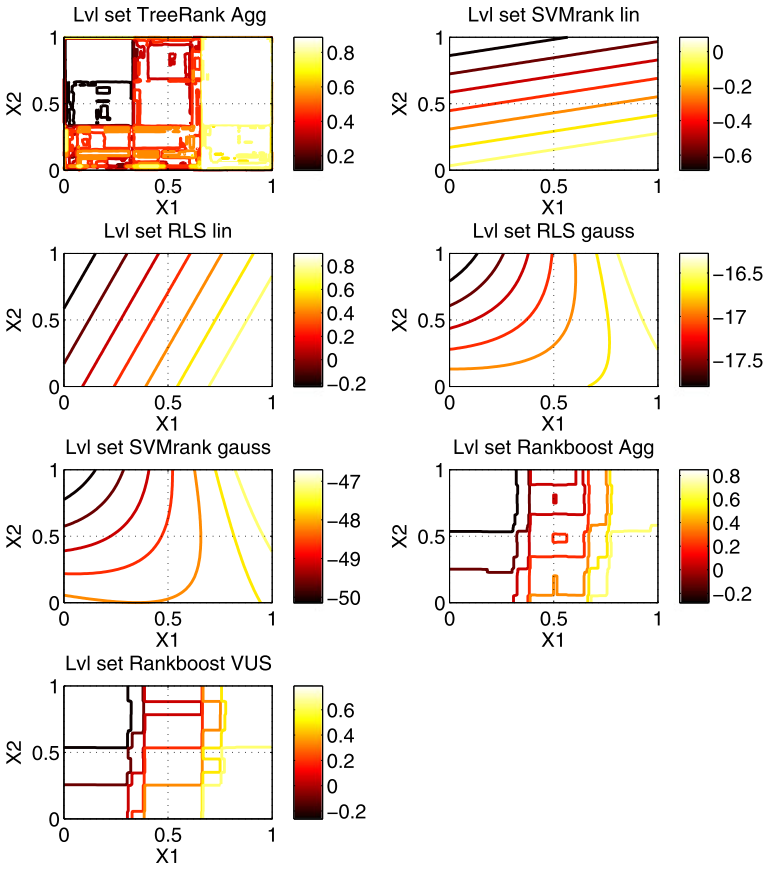


Fig. 6 Levels sets of the scoring functions "TreeRank Agg", "SVMrank lin", "RLScore lin", "RLScore gauss", "SVMrank gauss", "RankBoostVUS", "RankBoost Agg"

Table 4 Comparison of the VUS test—"Cardiotocography" experiment

Method	VUS test
TreeRank 1v2	0.2357
TreeRank 2v3	0.3314
TreeRank 1v3	0.6932
TreeRank Agg	0.8141
RankBoostVUS	0.8346
RankBoost Agg	0.8959
SVMrank lin	0.7202
SVMrank gauss	0.7856
RLScore lin	0.7652
RLScore gauss	0.7829

depending on the degree of anomaly observed: 1 for "normal", 2 for "suspect" and 3 for "pathologic".

We have split the data set into a training sample $\mathcal{D}_e$ and a test sample $\mathcal{D}_t$ of same sizes: scoring functions have been built based on the sample $\mathcal{D}_e$ and next tested on the sample $\mathcal{D}_t$ (*i.e.* we have computed the empirical versions of the ROC and VUS criteria based on $\mathcal{D}_t$). In this experiment, parameters have been selected by cross-validation: the scoring rule RankBoostVUS is based on 300 stumps and the bipartite rules produced by RankBoost are based on 100 stumps, the intercept involved in SVM ranklin is $C = (0.001)$, while SVMrank gauss, RLScore lin and RLScore gauss have been obtained with the respective parameters $(C, \gamma) = (0.001, 0.0001)$, $bias = 1$ and $(bias, \gamma) = (1, 0.001)$. Performance results are reported in Table 4 and the ROC surfaces test are plotted in Fig. 7.

Discussion We observe that, in each of these experiments, Kendall aggregation clearly improves ranking accuracy, when measured in terms of VUS. In addition, looking at the standard deviation, we see that the aggregated scoring function is more stable. In terms of level sets, Kendall aggregation yielded more complex subsets and thus sharper results. Notice additionally that, as in the "Gaussian" experiment the level sets are linear, it is not surprising that the kernel methods outperform the tree-based ones in this situation. In contrast, for the "uniform" experiment, the tree-based methods performed much better than the others, the performance of TreeRank Agg is nearly optimal. Looking at the level sets (see Fig. 6), they seem to recover well their geometric structure. Observe also that Kendall aggregation of (bipartite) scoring functions produced by RankBoost has always lead to (slightly) better results than those obtained by a direct use of RankBoost on the 3-class population, with a computation time smaller by a factor 10 however. Finally, notice that, on the Cardiotocography data set, the Kendall aggregation approach based on RankBoost is the method that produced the scoring function with largest VUS test among the algorithms candidates. In particular, it provides the best discrimination for the bipartite subproblem "1 vs 2", the most difficult to solve apparently, in view of the ROC surfaces plotted in Fig. 7.

Psychometric data For completeness, we also carried out experiments based on four datasets with ordinal labels (ERA, ESL, LEV and SWD namely), considered in David (2008) (due to the rarity of such data, regression datasets are sometimes transformed into ordinal datasets to serve as benchmarkk, see Fürnkranz et al. 2009 and Huhn and Hüllermeier 2008). Because of the wide disparity between certain class sizes, data with certain

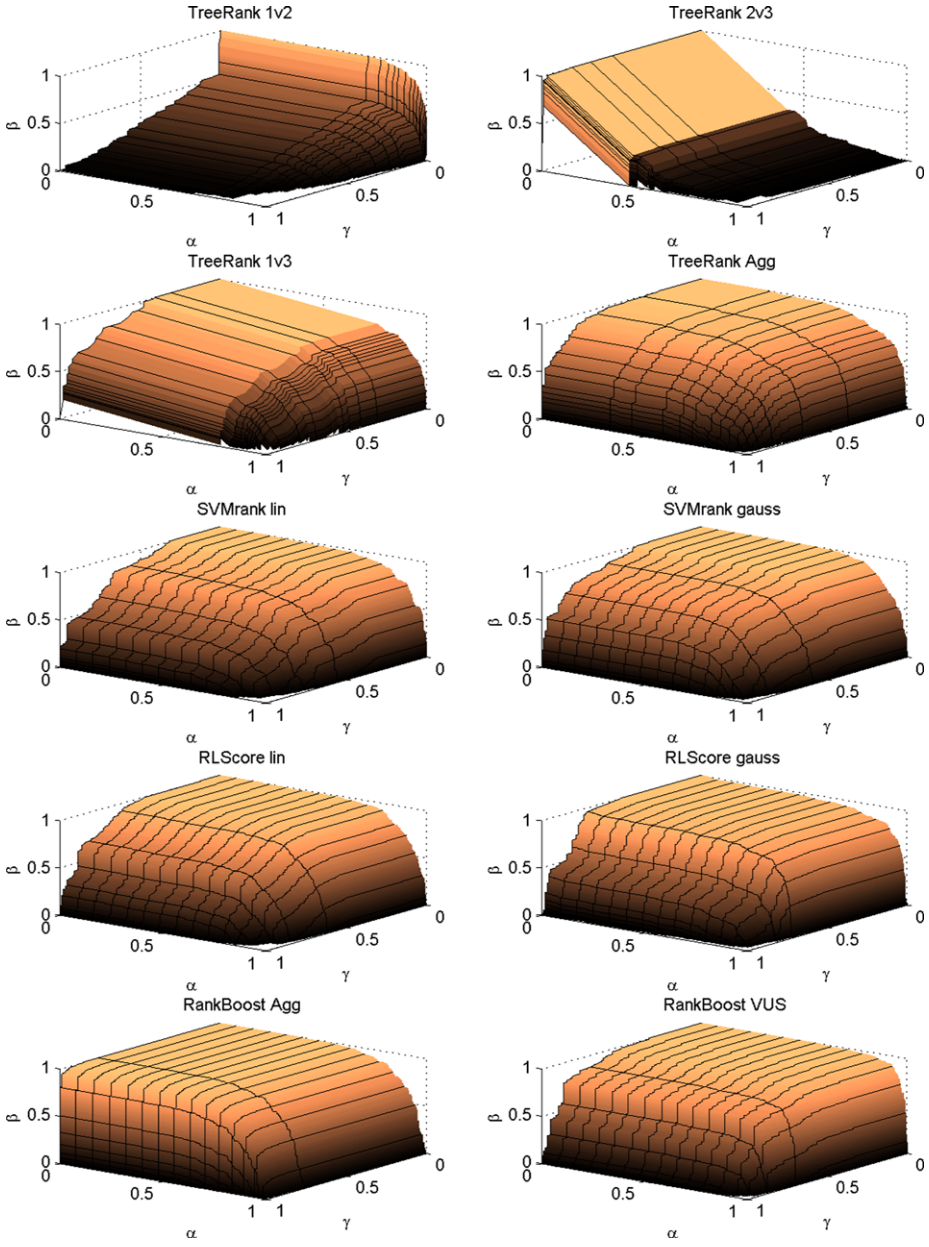


Fig. 7 ROC surfaces "test" of the scoring functions bipartite "TreeRank", "TreeRank Agg", "SVMrank lin", "RLScore lin", "RLScore gauss", "SVMrank gauss", "RankBoostVUS", "RankBoost Agg"

labels are ignored (in the ESL dataset for instance, the class "1" counts only two observations). On top of the ranking bipartite algorithms used previously, we also implemented the aggregation method based on the algorithm RANKING FOREST with 50 trees and a linear SVM with constant $C = 50$ as LEAFRANK procedure (see Clémençon et al. 2011b

Table 5 Comparison of the VUS test—"ERA" experiment

Method	class 1–7			class 1–9		
	VUS test	C-index	JPstat	VUS test	C-index	JPstat
TreeRank Agg	0.0068	0.7099	0.7292	0.0027	0.7330	0.8023
TreeRankF Agg	0.0074	0.7125	0.7326	0.0028	0.7359	0.8050
RankBoostVUS	0.0082	0.7141	0.7347	0.0029	0.7344	0.8065
RankBoost Agg	0.0077	0.7130	0.7331	0.0028	0.7329	0.8042
SVMrank lin	0.0088	0.7158	0.7359	0.0034	0.7380	0.8103
SVMrank gauss	0.0054	0.7033	0.7215	0.0020	0.7284	0.7969
RLScore lin	0.0090	0.7151	0.7354	0.0034	0.7386	0.8102
RLScore gauss	0.0080	0.7130	0.7331	0.0029	0.7339	0.8052

Table 6 Comparison of the VUS test—"ESL" experiment (class 3–7)

Method	VUS test	C-index	JPstat
TreeRank Agg	0.6209	0.9536	0.9551
TreeRankF Agg	0.6415	0.9588	0.9591
RankBoostVUS	0.5745	0.9496	0.9493
RankBoost Agg	0.5887	0.9513	0.9514
SVMrank lin	0.6337	0.9579	0.9583
SVMrank gauss	0.6074	0.9560	0.9544
RLScore lin	0.6387	0.9579	0.9590
RLScore gauss	0.6342	0.9568	0.9577

for more details), this is referred to as "TreeRankF Agg" in the tables. For each experiment, in addition to the VUS, we computed alternative ranking performance statistics over five replications of a five-fold cross validation: the C-INDEX (see Fürnkranz et al. 2009; Herbrich et al. 2000) and the JONCKHEERE-TERPSTRA STATISTIC (*JPstat* in abbreviated form, see Hand and Till 2001 and Higgins 2004). The results are reported in the Tables 5, 6, 7 and 8 (standard deviations are not indicated, their order of magnitude, 10^{-3} namely, being negligible).

We highlight the fact that the results we obtained with the approach promoted in this paper are quite comparable to those in Fürnkranz et al. (2009), they find a C-index of 0.7418 and a JPstat of 0.7265 in the case of ERA with nine classes as well as a C-index of 0.8660 and a JPstat of 0.8757 in the case of LEV with five classes. Contrary to the C-index and the JPstat for which all the values obtained are very close to each other, the VUS seems to reveal more contrast in the ranking performance. For instance, the aggregation procedure based on the TREERANK algorithm clearly outperforms the other competitors when considering the SWD dataset with classes 2–5, assessing the relevance of this approach in a situation where the dimension of the input space is not small (10 namely) and the population very skewed (the size of class "2" is very small compared to that of the others). Observe also that, in the other cases, the aggregation technique implemented using RANKING FOREST has performance very similar to the state-of-the-art: sometimes not considerably below (*cf.* ERA with 7 classes, ERA and LEV) sometimes slightly better (*cf.* ESL and SWD with 3 classes).

Table 7 Comparison of the VUS test—"LEV" experiment

Method	class 0–3			class 0–4		
	VUS test	C-index	JPstat	VUS test	C-index	JPstat
TreeRank Agg	0.4226	0.8347	0.8586	0.2932	0.8547	0.8758
TreeRankF Agg	0.4893	0.8617	0.8787	0.2995	0.8620	0.8761
RankBoostVUS	0.4842	0.8631	0.8773	0.2884	0.8637	0.8680
RankBoost Agg	0.4700	0.8570	0.8743	0.2761	0.8576	0.8703
SVMrank lin	0.4968	0.8668	0.8828	0.3124	0.8668	0.8753
SVMrank gauss	0.4870	0.8637	0.8783	0.2847	0.8638	0.8705
RLScore lin	0.4983	0.8668	0.8827	0.3122	0.8670	0.8751
RLScore gauss	0.4954	0.8639	0.8799	0.3215	0.8663	0.8797

Table 8 Comparison of the VUS test—"SWD" experiment

Method	class 2–5			class 3–5		
	VUS test	C-index	JPstat	VUS test	C-index	JPstat
TreeRank Agg	0.4221	0.8154	0.8674	0.5537	0.8072	0.8223
TreeRankF Agg	0.4169	0.8189	0.8659	0.5706	0.8150	0.8295
RankBoostVUS	0.3304	0.8141	0.8404	0.5619	0.8125	0.8280
RankBoost Agg	0.3562	0.8020	0.8498	0.5611	0.8127	0.8280
SVMrank lin	0.3278	0.8071	0.8369	0.5493	0.8083	0.8219
SVMrank gauss	0.3612	0.8140	0.8495	0.5599	0.8098	0.8238
RLScore lin	0.3316	0.8076	0.8386	0.5483	0.8078	0.8214
RLScore gauss	0.3680	0.8135	0.8518	0.5616	0.8123	0.8260

These empirical results only aim at illustrating the Kendall aggregation approach for K-partite ranking, the limited goal pursued here being to show how aggregation helps to improve results. Beyond the theoretical validity framework sketched in Sect. 3, since a variety of bipartite ranking algorithms have been proposed in the literature and dedicated libraries are readily available, one of the main advantages of the Kendall aggregation approach lies in the fact that it is very easy to implement, when applied to bipartite rules that are not too complex, so that the (approximate) median computation is feasible, see Sect. 3.2. A more complete and detailed empirical analysis of the merits and limitations of this procedure is currently the subject of ongoing work, where comparisons with competitors are carried out and computational issues are investigated at length, provided that more real datasets with ordinal labels can be obtained.

7 Conclusion

In this article, we have presented theoretical work on ranking data with ordinal labels. In the first part of the paper, the issue of optimality has been tackled. We have proposed a *monotonicity likelihood ratio condition* that guarantees the existence and unicity of an "optimal" preorder on the input space, in the sense that it is optimal for any bipartite ranking subproblem, considering all possible pairs of labels. In particular, the regression function is proved to

define an optimal ranking rule in this setting, highlighting the connection between K-partite ranking and ordinal regression. The second part is dedicated to describe a specific method for decomposing the multi-class ranking problem into a series of bipartite ranking tasks, as proposed in Fürnkranz et al. (2009). We have introduced a specific notion of *median scoring function* based on the (probabilistic) Kendall τ distance. We have next shown that the notion of ROC manifold/surface and its summary, the *volume under the* ROC *surface* (VUS), then provide quantitative criteria for evaluating ranking accuracy in the ordinal setup: under the afore mentioned monotonicity likelihood ratio condition, scoring functions whose ROC surface is as high as possible everywhere exactly coincide with those forming the optimal set (*i.e.* the set of scoring functions that are optimal for all bipartite subproblems, defined with no reference to the notions of ROC surface and VUS). Conversely, we have proved that the existence of a scoring function with such a dominating ROC surface implies that the monotonicity likelihood ratio condition is fulfilled. It is shown that the aggregation procedure leads to a consistent ranking rule, when applied to scoring functions that are, each, consistent for the bipartite ranking subproblem related to a specific pair of consecutive class distributions. This approach allows for extending the use of ranking algorithms originally designed for the bipartite situation to the ordinal multi-class context. It is illustrated by three numerical examples. Further experiments, based on more real datasets in particular, will be carried out in the future in order to determine precisely the situations in which this method is competitive, compared to alternative ranking techniques in the ordinal multi-class setup. In this respect, we underline that, so far, very few practical algorithms tailored for ROC graph optimization have been proposed in the literature. Whereas, as shown at length in Clémençon and Vayatis (2009b) and Clémençon et al. (2011a), partitioning techniques for AUC maximization, in the spirit of the CART method for classification, can be implemented in a very simple manner, by solving recursively cost-sensitive classification problems (with a local cost, depending on the data lying in the cell to be split), recursive VUS maximization remains a challenging issue, for which no simple interpretation is currently available. Hence, the number of possible strategies for direct optimization of the ranking criterion in the K-partite situation contrasts with that in the bipartite context and strongly advocates, for the moment, for considering techniques that transform multi-class ranking into a series of bipartite tasks, such as the method analyzed in this article.

Appendix A: Properties of the ROC surface

The next result summarizes several crucial properties of ROC surfaces. To the best of our knowledge, though expected, these properties have not been formulated in the literature. The technical proof straightforwardly relies on Proposition 17 in Clémençon and Vayatis (2009b) and the definition of the ROC surface given in Eq. (2), it is thus left to the reader.

Proposition 14 (PROPERTIES OF THE ROC SURFACE) *For any distributions $f_1(x)$, $f_2(x)$ and $f_3(x)$ on $\mathcal{X}$ and any scoring function $s \in \mathcal{S}$, the following properties hold.*

1. Intersections with the facets of the ROC space. *The intersection of the* ROC *surface $\{(\alpha, \mathrm{ROC}(s,\alpha),\gamma)\}$ with the plane of Eq. "$\alpha = 0$" coincides with the curve $\{(\beta, \mathrm{ROC}_{f_2,f_3}(s,\beta))\}$ up to the transform $(\beta,\gamma) \in [0,1]^2 \mapsto \psi(\beta,\gamma) = (1-\beta,\gamma)$, that with the plane of Eq. "$\beta = 0$" corresponds to the image of the curve $\{(\alpha, \mathrm{ROC}_{f_1,f_3}(s,\alpha))\}$ by the mapping $\psi(\alpha,\gamma)$ and that with the plane of Eq. "$\gamma = 0$" to the image of $\{(\alpha, \mathrm{ROC}_{f_1,f_2}(s,\alpha))\}$ by the transform $\psi(\alpha,\beta)$.*

2. Invariance. *For any strictly increasing function* $T : \mathbb{R} \cup \{+\infty\} \to \mathbb{R} \cup \{+\infty\}$, *we have, for all* $(\alpha, \gamma) \in [0, 1]^2$:

$$\mathrm{ROC}(T \circ s, \alpha, \gamma) = \mathrm{ROC}(s, \alpha, \gamma).$$

3. Concavity. *If the likelihood ratios* $dF_{s,2}/dF_{s,1}(u)$ *and* $dF_{s,3}/dF_{s,2}(u)$ *are both* (*strictly*) *increasing transforms of a certain function* $T(u)$, *then the* ROC *surface is* (*strictly*) *concave. In particular, if Assumption* 1 *is fulfilled, the surface* $\mathrm{ROC}^* \stackrel{def}{=} \mathrm{ROC}(s^*, ., .)$, *with* $s^* \in \mathcal{S}^*$, *is concave.*
4. Flat parts. *If the likelihood ratios* $dF_{s,2}/dF_{s,1}(u)$ *and* $dF_{s,3}/dF_{s,2}(u)$ *are simultaneously constant on some interval in the range of the scoring function* $s(x)$, *then the* ROC *surface will present a flat part* (*i.e. will be a part of a plane*) *on the corresponding domain. In addition, under the Assumption* 1, $(\alpha, \gamma) \mapsto \mathrm{ROC}^*(\alpha, \gamma)$ *is a linear function of* (α, γ) *on* $[\alpha_1, \alpha_2] \times [\gamma_1, \gamma_2] \subset \mathcal{I}_s$ *iff* $f_2/f_1(x)$ *and* $f_3/f_2(x)$ *are constant on the subsets*

$$\left\{x \in \mathcal{X} \mid Q\left(f_2/f_1(X), \alpha_2\right) \leq f_2/f_1(x) \leq Q\left(f_2/f_1(X), \alpha_1\right)\right\}$$

and

$$\left\{x \in \mathcal{X} \mid Q\left(f_3/f_2(X), \gamma_2\right) \leq f_3/f_2(x) \leq Q\left(f_3/f_2(X), \gamma_1\right)\right\}$$

respectively, denoting by $Q(Z, \alpha)$ *the quantile of order* $1 - \alpha$ *of any random variable* Z.
5. Differentiability. *Assume that the distributions* $f_1(x)$, $f_2(x)$ *and* $f_3(x)$ *are continuous. Then, the* ROC *surface of a scoring function* s *is differentiable if and only if the conditional distributions* $F_{s,1}(du)$, $F_{s,2}(du)$ *and* $F_{s,3}(du)$ *are continuous. In such a case, denoting by* $f_{s,1}$, $f_{s,2}$ *and* $f_{s,3}$ *the corresponding densities, we have in particular*: $\forall (\alpha, \gamma) \in \mathcal{I}_s$,

$$\frac{\partial}{\partial \alpha}\mathrm{ROC}(s, \alpha, \gamma) = -\frac{f_{s,2}}{f_{s,1}}\left(F_{s,1}^{-1}(\alpha)\right) \quad \textit{when } f_{s,1}\left(F_{s,1}^{-1}(\alpha)\right) > 0,$$

$$\frac{\partial}{\partial \gamma}\mathrm{ROC}(s, \alpha, \gamma) = -\frac{f_{s,2}}{f_{s,3}}\left(F_{s,3}^{-1}(1-\gamma)\right) \quad \textit{when } f_{s,3}\left(F_{s,3}^{-1}(1-\gamma)\right) > 0.$$

Preliminary results related to statistical estimation of the ROC surface of a fixed scoring function $s(x)$ can be found in Li and Zhou (2009), additional results related to the building of confidence regions in the ROC space $[0, 1]^3$ are established in Robbiano (2010).

Alternative ROC *graph* Another way of quantifying the ranking accuracy of a scoring function in the multi-class setting is to evaluate its ability to discriminate between X's conditional distributions given $Y \leq k$ and $Y > k$ respectively, which we denote $h_k(x)$ and $g_k(x)$, for $k \in \{1, \ldots, K-1\}$. This boils down to plot the graph of the mapping $\alpha \in [0, 1] \mapsto (\mathrm{ROC}_{h_1,g_1}(s, \alpha), \ldots, \mathrm{ROC}_{h_{K-1},g_{K-1}}(s, \alpha))$. It straightforwardly follows from the stipulated monotonicity hypothesis (*cf.* Assumption 1) that the curve related to $s^* \in \mathcal{S}^*$ dominates the curve of any other scoring function s in the coordinatewise sense: $\mathrm{ROC}_{h_k,g_k}(s, \alpha) \leq \mathrm{ROC}_{h_k,g_k}(s^*, \alpha)$ for all $\alpha \in [0, 1]$, $1 \leq k < K$. The likelihood ratio $g_k/h_k(X)$ is indeed a non decreasing function of $s^*(X)$, see Theorem 3.4.1 in Lehmann and Romano (2005) for instance. However, with such a functional representation of ranking performance, one loses an attractive advantage, the insensitivity to the class probabilities p_k. Indeed, the distributions $h_k(x)$ and $g_k(x)$ depend on the latter, they can be expressed as $\sum_{l \leq k} p_l f_l(x)/(\sum_{l \leq k} p_l)$ and $\sum_{l > k} p_l f_l(x)/(\sum_{l > k} p_l)$ respectively.

On the ROC *surface of a classification rule* We point out that, with the convention previously introduced, the ROC surface of a classifier $C : \mathcal{X} \to \{1, 2, 3\}$ is the polyhedron with vertices $(0, 0, 1)$, $(0, \alpha_{2,1}, 1 - \alpha_{3,1})$, $(0, 1 - \alpha_{2,3}, \alpha_{3,3})$, $(0, 1, 0)$, $(\alpha_{1,1}, 0, 1 - \alpha_{3,1})$, $(\alpha_{1,1}, \alpha_{2,2}, \alpha_{3,3})$, $(\alpha_{1,1}, 1 - \alpha_{2,1}, 0)$, $(1 - \alpha_{1,3}, 0, \alpha_{3,3})$, $(1 - \alpha_{1,3}, \alpha_{2,3}, 0)$ and $(1, 0, 0)$, where $\alpha_{k,l} = \mathbb{P}\{C(X) = l \mid Y = k\}$. We underline that the confusion matrix $\mathcal{M}(C) = \{\alpha_{k,l}\}$ can be fully recovered from this geometric solid, which is actually a decahedron when the matrix $\mathcal{M}(C)$ has no null entry. Observe finally that this graphic representation of $\mathcal{M}(C)$ differs from that which derives from the multi-class notion of ROC analysis proposed in Ferri et al. (2003). In the latter case, the ROC space is defined as $[0, 1]^6$ and $\mathcal{M}(C)$ is represented by the point with coordinates $(\alpha_{1,2}, \alpha_{1,3}, \alpha_{2,1}, \alpha_{2,3}, \alpha_{3,1}, \alpha_{3,2})$. Notice incidentally that the latter concept of ROC analysis is more general in the sense that it permits to visualize the performance of $K(K-1)/2$ classifiers involved in a *one-versus-one* classification method.

Appendix B: Technical proofs

B.1 Proof of Proposition 1

The assertions (3) $\Rightarrow$ (2), (2) $\Rightarrow$ (4) and (2) $\Rightarrow$ (5) are straightforward.

(1) $\Rightarrow$ (3) Recall that $\eta(x) = \sum_{k=1}^{K} k \cdot \eta_k(x)$. Our goal is to establish that: $\forall (x, x') \in \mathcal{X}^2$,

$$\Phi_{k,l}(x) < \Phi_{k,l}\big(x'\big) \Rightarrow \eta(x) < \eta\big(x'\big).$$

The proof is based on the next lemma.

Lemma 2 *Suppose Assumption 1 is satisfied. Let* $(x, x') \in \mathcal{X}^2$. *If there exists* $1 \leq l < k \leq K$ *such that* $0 < \Phi_{k,l}(x) < \Phi_{k,l}(x')$, *then for all* $j \in \{1, \ldots, K\}$, *we have*

$$\sum_{i=j}^{K} \eta_i(x) \leq \sum_{i=j}^{K} \eta_i\big(x'\big). \tag{6}$$

Additionally, a strict version of inequality (6) *holds true when* $j = l + 1$.

Proof Let $(x, x') \in \mathcal{X}^2$ and $1 \leq l < k \leq K$ be such that $\Phi_{k,l}(x) < \Phi_{k,l}(x')$

Combining $\Phi_{k,l}(x) = \frac{p_l \eta_k(x)}{p_k \eta_l(x)}$ and $\eta_l(x) = 1 - \sum_{i \neq l} \eta_i(x)$, we clearly have

$$\eta_k(x) - \eta_k(x) \sum_{i \neq l} \eta_i\big(x'\big) < \eta_k\big(x'\big) - \eta_k\big(x'\big) \sum_{i \neq l} \eta_i(x),$$

and, by virtue of Assumption 1, for $1 \leq j \leq m \leq K$:

$$\eta_m(x) \leq \eta_m\big(x'\big) + \sum_{i < j-1} \big\{\eta_m(x)\eta_i\big(x'\big) - \eta_m\big(x'\big)\eta_i(x)\big\} \tag{7}$$

$$+ \sum_{i > j-1} \big\{\eta_m(x)\eta_i\big(x'\big) - \eta_m\big(x'\big)\eta_i(x)\big\},$$

$$\leq \eta_m\big(x'\big) + \sum_{i > j-1} \big\{\eta_m(x)\eta_i\big(x'\big) - \eta_m\big(x'\big)\eta_i(x)\big\}. \tag{8}$$

Summing up, term-by-term, inequalities (7) for $m = j, \ldots, K$, one gets that

$$\sum_{m=j}^{K} \eta_m(x) \leq \sum_{m=j}^{K} \eta_m(x') + \sum_{m=j}^{K} \sum_{i=j}^{K} \{\eta_m(x)\eta_i(x') - \eta_m(x')\eta_i(x)\}.$$

The proof is finished by noticing that the sum on the right hand side of the inequality above is equal to 0. □

The desired result is established by summing up the inequalities (6) stated in Lemma 2 for $j = 1, \ldots, K$.

(4) ⇒ (2) Using the fact that $\Phi_{k,l}(x) = \prod_{j=l}^{k-1} \Phi_{j+1,j}(x)$, immediatly gives the result.

(5) ⇒ (1) We call $\Psi_{k,l}$ the nondecreasing function such that $\Phi_{k,l}(x) = \Psi_{k,l}(s^*(x))$. Let $(k,l) \in \{1, \ldots, K\}^2$ s.t. $l < k$ and x, x' in $\mathcal{X}_k \cap \mathcal{X}_l$. Suppose that $\Phi_{k+1,k}(x) < \Phi_{k+1,k}(x')$. The functions $\Psi_{k,l}^{-1}$ are nondecreasing, just like the $\Psi_{k,l}$'s. The equality $\Phi_{l+1,l}(x) = \Psi_{l+1,l}(\Psi_{k+1,k}^{-1} \circ \Phi_{k+1,k}(x))$ on $\mathcal{X}_k \cap \mathcal{X}_l$ leads to the result.

B.2 Proof of Proposition 2

Notice first that it is actually sufficient to prove that $\mathcal{X}_{k-1} \cap \mathcal{X}_{k+1} \subset \mathcal{X}_k$ for all $k \in \{2, \ldots, K-1\}$. Let $1 < k < K$ and suppose that $\mathcal{X}_{k-1} \cap \mathcal{X}_k \neq \emptyset$ (the inclusion is immediate otherwise). Consider $x \in \mathcal{X}_{k-1} \cap \bar{\mathcal{X}}_k \cap \mathcal{X}_{k+1}$, where $\bar{\mathcal{X}}_k = \mathcal{X} \setminus \mathcal{X}_k$. We thus have: $\Phi_{k,k-1}(x) = 0$ and $\Phi_{k+1,k}(x) = +\infty$. Hence, for any $x' \in \mathcal{X}_k$, we have: $0 = \Phi_{k,k-1}(x) \leq \Phi_{k,k-1}(x')$ and $\Phi_{k+1,k}(x') \leq \Phi_{k+1,k}(x) = +\infty$. Assumption 1 implies that both inequalities are actually equalities, which is in contradiction with the fact that $x' \in \mathcal{X}_k$.

B.3 Proof of Proposition 3

Under Assumption 1, the regression function η is an optimal scoring function (see Theorem 1(3)). Using the fact that $\phi_{k+1,k} \in \mathcal{S}_{k+1,k}^*$ combined with Theorem 1(4), we obtain $\tau(s_k^*, \eta) = 1$ for $k = 1, \ldots, K-1$. As $\eta \in \mathcal{S}_1$, it achieves the maximum over the class $\mathcal{S}_1$, yielding (2). Hence, for any median scoring rule $\bar{s}$, we have $\tau(s_k^*, \bar{s}) = 1$ for $k \in \{1, \ldots, K-1\}$, i.e. $\bar{s} \in \mathcal{S}_{k+1,k}^*$ for $k \in \{1, \ldots, K-1\}$, and thus $\bar{s} \in \mathcal{S}^*$.

B.4 Proof of Proposition 4

This results from the change of parameters: $\alpha = F_{s,1}(t_1)$ and $\gamma = 1 - F_{s,3}(t_2)$.

B.5 Proof of Lemma 1

(1) ⇒ (2) If $\mathrm{ROC}(s, \alpha, \gamma) > 0$ using Proposition 4, we get $t_1 < t_2$, $\alpha = F_{s,1}(t_1)$ and $\gamma = 1 - F_{s,3}(t_2)$. Using the definition of the ROC curve and $t_1 < t_2$, we have $\mathrm{ROC}_{f_1,f_3}(s, 1-\gamma) = 1 - F_{s,3}(t_1) > 1 - F_{s,3}(t_2) = \gamma$. (2) ⇒ (1) If $\mathrm{ROC}_{f_1,f_3}(s, 1-\gamma) > \gamma$ then $1 - F_{s,3}(t_1) > 1 - F_{s,3}(t_2)$ so $t_1 < t_2$, and $F_{s,2}(t_2) - F_{s,2}(t_1) > 0$. This yields the desired result.

B.6 Proof of Theorem 1

Let $(s, s^*) \in \mathcal{S} \times \mathcal{S}^*$. Since, in particular, the scoring function s^* belongs to the set $\mathcal{S}_{1,3}^*$, we have $\mathrm{ROC}_{F_1,F_3}(s^*, 1-\alpha) \geq \mathrm{ROC}_{F_1,F_3}(s, 1-\alpha)$ for all $\alpha \in [0,1]$. Hence, as the desired bound obviously holds true on the set $\{(\alpha, \gamma)\colon \gamma > \mathrm{ROC}_{F_1,F_3}(s^*, 1-\alpha)\} \subset$

$\{(\alpha,\gamma)\colon \gamma > \mathrm{ROC}_{F_1,F_3}(s,1-\alpha)\}$, we place ourselves on the complementary set $\{(\alpha,\gamma)\colon \gamma \leq \mathrm{ROC}_{F_1,F_3}(s^*,1-\alpha)\}$, on which we have

$$\mathrm{ROC}\big(s^*,\alpha,\gamma\big) - \mathrm{ROC}(s,\alpha,\gamma) \leq \big(\mathrm{ROC}_{f_1,f_2}\big(s^*,1-\alpha\big) - \mathrm{ROC}_{f_1,f_2}(s,1-\alpha)\big) + \big(\mathrm{ROC}_{f_3,f_2}(s,\gamma) - \mathrm{ROC}_{f_3,f_2}\big(s^*,\gamma\big)\big).$$

The terms on the right hand side of the equation are both nonnegative, since s^* lies in $\mathcal{S}^*_{1,2}$ and $\mathcal{S}^*_{3,2}$ respectively (observing that, whatever the two distributions H and G on $\mathbb{R}$ and for any $s\in\mathcal{S}$ and $(\alpha,\beta)\in[0,1]^2$, we have: $\mathrm{ROC}_{H,G}(s,\alpha)\leq\beta \Leftrightarrow \alpha \leq \mathrm{ROC}_{G,H}(s,\beta)$). The first part of the result is thus established.

Suppose that there exists $s^*\in\mathcal{S}$ such that, for any $s\in\mathcal{S}$, we have: $\forall(\alpha,\gamma)\in[0,1]^2$,

$$\mathrm{ROC}\big(s^*,\alpha,\gamma\big) \geq \mathrm{ROC}(s,\alpha,\gamma). \tag{9}$$

Observe that, if $\gamma > \mathrm{ROC}_{f_1,f_3}(s^*,1-\alpha)$, this implies that $\gamma > \mathrm{ROC}_{f_1,f_3}(s,1-\alpha)$, whatever (α,γ). It then follows that $s^*\in\mathcal{S}^*_{1,3}$. Now the fact that s^* belongs to $\mathcal{S}^*_{1,2}$ (respectively, to $\mathcal{S}^*_{1,3}$) straightforwardly result from Eq. (9) with $\beta=0$ (respectively, with $\alpha=1$).

B.7 Proof of Proposition 5

We denote by $\bar{E}=\mathcal{X}\setminus E$ the complementary set of any subset $E\subset\mathcal{X}$ and set $m_1(x)=\mathbb{I}\{x\in\bar{R}^{*(1)}_\alpha\}-\mathbb{I}\{x\in\bar{R}^{(1)}_{s,\alpha}\}$ and $m_3(x)=\mathbb{I}\{x\in R^{*(3)}_{1-\gamma}\}-\mathbb{I}\{x\in R^{(3)}_{s,1-\gamma}\}$ for $\alpha\in[0,1]$. On the set $\{(\alpha,\gamma)\colon \gamma\leq\mathrm{ROC}_{f_1,f_3}(s^*,1-\alpha)\}$, we may then write:

$$\mathrm{ROC}\big(s^*,\alpha,\gamma\big) - \mathrm{ROC}(s,\alpha,\gamma) \leq -\mathbb{E}\big[m_1(X)|Y=2\big] - \mathbb{E}\big[m_3(X)|Y=2\big].$$

Considering the first ROC curve deficit, we have:

$$-\mathbb{E}\big[m_1(X)|Y=2\big] = -\frac{p_1}{p_2}\mathbb{E}\left[m_1(X)\frac{\eta_2(X)}{\eta_1(X)}\,\middle|\, Y=1\right].$$

Then we add and subtract $\frac{\eta_3(x)}{\eta_1(x)} - \frac{1-Q^{(1)}(\eta_1,\alpha)}{Q^{(1)}(\eta_1,\alpha)}$, this leads to:

$$-\mathbb{E}\big[m_1(X)|Y=2\big] = -\frac{p_1}{p_2}\mathbb{E}\left[m_1(X)\left(\frac{\eta_2(X)+\eta_3(X)}{\eta_1(X)} + \frac{1-Q^{(1)}(\eta_1,\alpha)}{Q^{(1)}(\eta_1,\alpha)}\right)\middle| Y=1\right] + \frac{p_1}{p_2}\mathbb{E}\left[m_1(x)\frac{\eta_3(X)}{\eta_1(X)}\middle| Y=1\right].$$

By definition of s^*, the second term on the right hand side of the equation above is equal to

$$\frac{p_3}{p_2}\mathbb{E}\big[m_1(X)|Y=3\big] = \mathrm{ROC}_{f_1,f_3}(s,1-\alpha) - \mathrm{ROC}_{f_1,f_3}\big(s^*,1-\alpha\big),$$

while, for the first term, by removing the conditioning with respect to $Y=1$ and using then the definition of $Q^{(1)}(\eta_1,\alpha)$, we get:

$$\frac{1}{p_2 Q^{(1)}(\eta_1,\alpha)}\mathbb{E}\big[m_1(X)\big(\eta_1(X)-Q^{(1)}(\eta_1,\alpha)\big)\big] = \frac{1}{p_2}\mathbb{E}\big[\big|\eta_1(X)-Q^{(1)}(\eta_1,\alpha)\big|m_1(X)\big].$$

The first part of the desired bound follows from $A\Delta B=\bar{A}\Delta\bar{B}$. The other ROC curve difference can be handled the same way. This leads to the desired result.

B.8 Proof of Proposition 6

We have:

$$\begin{aligned}\mathrm{VUS}(s) &= \int_0^1 \int_0^1 \mathrm{ROC}(s,\alpha,\gamma)\,\mathrm{d}\alpha\,\mathrm{d}\gamma = \int_0^1 \int_0^{1-\gamma} (1-\alpha-\gamma)\,\mathrm{d}\alpha\,\mathrm{d}\gamma \\ &= \frac{1}{2}\int_0^1 (1-\gamma)^2\,\mathrm{d}\alpha\,\mathrm{d}\gamma = 1/6,\end{aligned}$$

which establishes the first assertion, while the second one results from:

$$\mathrm{VUS}(s) = \int_0^1 \int_0^1 \mathrm{d}\alpha\,\mathrm{d}\gamma = 1.$$

B.9 Proof of Proposition 8

The result simply follows from integration over $(\alpha,\gamma) \in [0,1]^2$ of the inequality stated in Theorem 1.

B.10 Proof of Proposition 9 (sketch of)

This result simply derives from Proposition 4 in Clémençon and Vayatis (2009b), applied to the bipartite ranking subproblems related to the pairs $(k,\ k+1)$, with $1 \le k < K$.

B.11 Proof of Theorem 2

Let $(s,s^*) \in \mathcal{S} \times \mathcal{S}^*$. Notice that, as $s^* \in \mathcal{S}^*_{1,3}$, we have $\{(\alpha,\gamma)\colon \gamma \le \mathrm{ROC}_{f_1,f_3}(s,1-\alpha)\} \subset \{(\alpha,\gamma)\colon \gamma \le \mathrm{ROC}_{f_1,f_3}(s^*,1-\alpha)\}$, so that

$$\begin{aligned}\mathrm{ROC}^*(\alpha,\gamma) - \mathrm{ROC}(s,\alpha,\gamma) &\le \big\{\mathrm{ROC}_{f_1,f_2}\big(s^*,1-\alpha\big) - \mathrm{ROC}_{f_3,f_2}\big(s^*,\gamma\big) \\ &\quad - \big(\mathrm{ROC}_{f_1,f_2}(s,1-\alpha) - \mathrm{ROC}_{f_3,f_2}(s,\gamma)\big)_+\big\} \\ &\quad \times \mathbb{I}\big\{\gamma \le \mathrm{ROC}^*_{f_1,f_3}(1-\alpha)\big\} \\ &\le \big\{\mathrm{ROC}_{f_1,f_2}\big(s^*,1-\alpha\big) - \mathrm{ROC}_{f_3,f_2}\big(s^*,\gamma\big) \\ &\quad - \mathrm{ROC}_{f_1,f_2}(s,1-\alpha) - \mathrm{ROC}_{f_3,f_2}(s,\gamma)\big\} \\ &\quad \times \mathbb{I}\big\{\gamma \le \mathrm{ROC}^*_{f_1,f_3}(1-\alpha)\big\} \\ &\le \big(\mathrm{ROC}_{f_1,f_2}\big(s^*,1-\alpha\big) - \mathrm{ROC}_{f_1,f_2}(s,1-\alpha)\big) \\ &\quad - \big(\mathrm{ROC}_{f_3,f_2}\big(s^*,\gamma\big) - \mathrm{ROC}_{f_3,f_2}(s,\gamma)\big).\end{aligned}$$

Integrating over $(\alpha,\gamma) \in [0,1]^2$ then yields the desired bound, using the fact that, for any $s \in \mathcal{S}_0$, $\int_{\gamma=0}^1 \mathrm{ROC}_{f_3,f_2}(s,\gamma)\,\mathrm{d}\gamma = 1 - \mathrm{AUC}_{f_2,f_3}(s)$.

B.12 Proof of Proposition 10 (sketch of)

Observe that this result corresponds to Corollary 8 in Clémençon et al. (2008), applied to the regression function obtained when conditioning upon the event $Y \in \{k, k+1\}$, $1 \le k < K$. We refer to the argument of the latter.

B.13 Proof of Proposition 11

Recall that $\tau(s_1, s_2) = 1 - 2d_\tau(s_1, s_2)$, where $d_\tau(s_1, s_2)$ is given by:

$$\mathbb{P}\big\{\big(s_1(X) - s_1\big(X'\big)\big) \cdot \big(s_2(X) - s_2\big(X'\big)\big) < 0\big\} + \frac{1}{2}\mathbb{P}\big\{s_1(X) = s_1\big(X'\big), s_2(X) \neq s_2\big(X'\big)\big\}$$
$$+ \frac{1}{2}\mathbb{P}\big\{s_1(X) \neq s_1\big(X'\big), s_2(X) = s_2\big(X'\big)\big\}.$$

Observe first that, for all $s \in \mathcal{S}_0$, $\mathrm{AUC}_{f_1,f_2}(s)$ may be written as:

$$\mathbb{P}\big\{\big(s(X) - s\big(X'\big)\big) \cdot \big(Y - Y'\big) > 0\big\}/\big(2p(1-p)\big) + \mathbb{P}\big\{s(X) = s\big(X'\big),\ Y \neq Y'\big\}/\big(4p(1-p)\big).$$

Notice also that, using Jensen's inequality, one easily obtain that the quantity $2p(1-p)|\mathrm{AUC}_{f_1,f_2}(s_1) - \mathrm{AUC}_{f_1,f_2}(s_2)|$ is bounded by the expectation of the random variable

$$\mathbb{I}\big\{\big(s_1(X) - s_1\big(X'\big)\big) \cdot \big(s_2(X) - s_2\big(X'\big)\big) < 0\big\} + \frac{1}{2}\mathbb{I}\big\{s_1(X) = s_1\big(X'\big)\big\} \cdot \mathbb{I}\big\{s_2(X) \neq s_2\big(X'\big)\big\}$$
$$+ \frac{1}{2}\mathbb{I}\big\{s_1(X) \neq s_1\big(X'\big)\big\} \cdot \mathbb{I}\big\{s_2(X) = s_2\big(X'\big)\big\},$$

which is equal to $d_\tau(s_1, s_2) = (1 - \tau(s_1, s_2))/2$. This proves the assertion.

B.14 Proof of Proposition 12

Set $\Gamma_s = \{(x, x') \in \mathcal{X}^2 \colon (\zeta(x) - \zeta(x'))(s(x) - s(x')) < 0\}$. We have, for all real valued scoring functions $(s, s^*) \in \mathcal{S} \times \mathcal{S}^*_{1,2}$:

$$d_\tau\big(s, s^*\big) \leq \mathbb{P}\big\{\big(X, X'\big) \in \Gamma_s\big\} + \frac{1}{2}\mathbb{P}\big\{s(X) = s\big(X'\big)\big\}.$$

Recall also that

$$2p(1-p)\big(\mathrm{AUC}^*_{f_1,f_2} - \mathrm{AUC}_{f_1,f_2}(s)\big) = \mathbb{E}\big[\big|\zeta(X) - \zeta\big(X'\big)\big|\mathbb{I}\big\{\big(X, X'\big) \in \Gamma_s\big\}\big]$$
$$+ \mathbb{P}\big\{s(X) = s\big(X'\big), \big(Y, Y'\big) = (-1, +1)\big\},$$

see Example 1 in Clémençon et al. (2008) for instance.

Observe that Hölder inequality combined with the noise condition shows that the quantity $\mathbb{E}\,[\mathbb{I}\{(X, X') \in \Gamma_s\}]$ is bounded by

$$\mathbb{E}\big[\big|\zeta(X) - \zeta\big(X'\big)\big| \cdot \mathbb{I}\big\{\big(X, X'\big) \in \Gamma_s\big\}\big]^{a/(1+a)} c^{1/(1+a)}.$$

In addition, we have

$$\mathbb{P}\big\{s(X) = s\big(X'\big), \big(Y, Y'\big) = (-1, +1)\big\}$$
$$= \frac{1}{2}\mathbb{E}\big[\mathbb{I}\big\{s(X) = s\big(X'\big)\big\} \cdot \big(\zeta(X) + \zeta\big(X'\big) - 2\zeta(X)\zeta\big(X'\big)\big)\big],$$

and the upper bound can be easily seen as larger than $\mathbb{E}[\mathbb{I}\{s(X) = s(X')\} \cdot |\zeta(X) - \zeta(X')]/2$. Therefore, using the same Hölder argument as above, we obtain that

$$\mathbb{P}\big\{s(X) = s\big(X'\big)\big\} \leq \big(\mathbb{E}\big[\big|\zeta(X) - \zeta\big(X'\big)\big| \cdot \mathbb{I}\big\{s(X) = s\big(X'\big)\big\}\big]\big)^{a/(1+a)} \times c^{1/(1+a)}.$$

Combining the bounds above, the concavity of $t \mapsto t^{a/(1+a)}$ permits to finish the proof.

B.15 Proof of Theorem 3

Let $(s_n^{(1)}, s_n^{(2)})$ be a sequence of real-valued scoring functions in $\mathcal{S}_1$ such that, as $n \to \infty$, $\mathrm{AUC}_{f_1,f_2}(s_n^{(1)}) \to \mathrm{AUC}^*_{f_1,f_2}$ and $\mathrm{AUC}_{f_2,f_3}(s_n^{(2)}) \to \mathrm{AUC}^*_{f_2,f_3}$. Here we consider the following consensus measure: $\forall s \in \mathcal{S}_1$,

$$\Delta_n(s) = d_\tau\big(s, s_n^{(1)}\big) + d_\tau\big(s, s_n^{(2)}\big).$$

Let $s^* \in \mathcal{S}_1 \cap \mathcal{S}^*$. Denote by $d_{\tau_{1,2}}$ the Kendall tau distance when $X \sim (p_1/(1-p_3))F_1 + (p_2/(1-p_3))F_2$. Proposition 11, combined with the triangular inequality applied to the pseudo-distance $d_{\tau_{1,2}}$, implies that

$$\begin{aligned}
\mathrm{AUC}^*_{f_1,f_2} - \mathrm{AUC}_{f_1,f_2}(\bar{s}_n) &\leq \frac{d_{\tau_{1,2}}(s^*, \bar{s}_n)}{p_1 p_2/(1-p_3)^2} \\
&\leq \frac{d_{\tau_{1,2}}(\bar{s}_n^{(1)}, \bar{s}_n) + d_{\tau_{1,2}}(s^*, \bar{s}_n^{(1)})}{p_1 p_2/(1-p_3)^2} \\
&\leq \frac{d_{\tau_{1,2}}(s^*, \bar{s}_n^{(1)})}{p_1 p_2/(1-p_3)^2} + \frac{d_\tau(\bar{s}_n^{(1)}, \bar{s}_n)}{p_1 p_2}.
\end{aligned}$$

The desired result follows from Proposition 12 combined with the AUC-consistency assumptions.

B.16 Proof of Proposition 13

By virtue of Theorem 2, we have:

$$\mathrm{VUS}^* - \mathrm{VUS}(\widehat{\eta}) \leq \big(\mathrm{AUC}^*_{f_1,f_2} - \mathrm{AUC}_{f_1,f_2}(\widehat{\eta})\big) + \big(\mathrm{AUC}^*_{f_2,f_3} - \mathrm{AUC}_{f_2,f_3}(\widehat{\eta})\big).$$

Considering the first term on the right hand side of the equation above, we have:

$$\mathrm{AUC}^*_{f_1,f_2} - \mathrm{AUC}_{f_1,f_2}(\widehat{\eta}) = \frac{1}{2p_1p_2}\mathbb{E}\big[\big|\eta_1(X)\eta_2(X') - \eta_1(X')\eta_2(X)\big| \cdot \mathbb{I}\big\{(X,X') \in \Gamma\big\}\big],$$

where

$$\Gamma = \big\{(x,x') \in \mathcal{X}^2 \colon \big(\eta(x) - \eta(x')\big)\big(\widehat{\eta}(x) - \widehat{\eta}(x')\big) < 0\big\}.$$

By using the triangular inequality and Lemma 2, one may establish that: $\forall (x,x') \in \mathcal{X}^2$, $\forall i \in \{1,2,3\}$,

$$\big|\eta_i(x) - \eta_i(x')\big| < \big|\eta(x) - \eta(x')\big|.$$

Then, we get:

$$\mathrm{AUC}^*_{f_1,f_2} - \mathrm{AUC}_{f_1,f_2}(\widehat{\eta}) \leq \frac{1}{2p_1p_2}\mathbb{E}\big[\big|\eta(X) - \eta(X')\big|\mathbb{I}\big\{(X,X') \in \Gamma\big\}\big].$$

But, one may easily check that, if $(x,x') \in \Gamma$, then

$$\big|\eta(x) - \eta(x')\big| \leq \big|\eta(x) - \widehat{\eta}(x)\big| + \big|\eta(x') - \widehat{\eta}(x')\big|.$$

As the same argument can be applied to the second AUC difference, this gives the desired result.

References

Agarwal, S. (2008). Generalization bounds for some ordinal regression algorithms. In *Proceedings of the 19th international conference on algorithmic learning theory, ALT '08* (pp. 7–21). Berlin: Springer.

Agarwal, S., Graepel, T., Herbrich, R., Har-Peled, S., & Roth, D. (2005). Generalization bounds for the area under the ROC curve. *Journal of Machine Learning Research*, *6*, 393–425.

Allwein, E., Schapire, R., & Singer, Y. (2001). Reducing multiclass to binary: a unifying approach for margin classifiers. *Journal of Machine Learning Research*, *1*, 113–141.

Audibert, J., & Tsybakov, A. (2007). Fast learning rates for plug-in classifiers. *The Annals of Statistics*, *35*, 608–633.

Barthélemy, J., Guénoche, A., & Hudry, O. (1989). Median linear orders: heuristics and a branch and bound algorithm. *European Journal of Operational Research*, *42*(3), 313–325.

Baskiotis, N., Clémençon, S., Depecker, M., & Vayatis, N. (2010). Treerank: an R package for bipartite ranking. In *Proceedings of SMDTA 2010—stochastic modeling techniques and data analysis international conference*.

Beygelzimer, A., Dani, V., Hayes, T., Langford, J., & Zadrozny, B. (2005a). Error limiting reductions between classification tasks. In *Machine learning, proceedings of the twenty-second international conference (ICML 2005)* (pp. 49–56).

Beygelzimer, A., Langford, J., & Zadrozny, B. (2005b). Weighted one against all. In *Proceedings of the 20th national conference on artificial intelligence, AAAI '05* (Vol. 2, pp. 720–725).

Charon, I., & Hudry, O. (1998). Lamarckian genetic algorithms applied to the aggregation of preferences. *Annals of Operations Research*, *80*, 281–297.

Clémençon, S., & Robbiano, S. (2011). Minimax learning rates for bipartite ranking and plug-in rules. In *Proceedings of the 28th international conference on machine learning, ICML'11* (pp. 441–448).

Clémençon, S., & Vayatis, N. (2009a). On partitioning rules for bipartite ranking. *Journal of Machine Learning Research*, *5*, 97–104.

Clémençon, S., & Vayatis, N. (2009b). Tree-based ranking methods. *IEEE Transactions on Information Theory*, *55*(9), 4316–4336.

Clémençon, S., & Vayatis, N. (2009c). Adaptive estimation of the optimal ROC curve and a bipartite ranking algorithm. In *Proceedings of the 20th international conference on algorithmic learning theory, ALT '09* (pp. 216–231).

Clémençon, S., & Vayatis, N. (2010). Overlaying classifiers: a practical approach to optimal scoring. *Constructive Approximation*, *32*(3), 619–648.

Clémençon, S., Lugosi, G., & Vayatis, N. (2008). Ranking and empirical risk minimization of U-statistics. *The Annals of Statistics*, *36*(2), 844–874.

Clémençon, S., Depecker, M., & Vayatis, N. (2011a). Adaptive partitioning schemes for bipartite ranking. *Machine Learning*, *43*(1), 31–69.

Clémençon, S., Depecker, M., & Vayatis, N. (2011b). Avancées récentes dans le domaine de l'apprentissage statistique d'ordonnancements. *Revue d'Intelligence Artificielle*, *25*(3), 345–368.

David, A. B. (2008). Ordinal real-world data sets repository.

Debnath, R., Takahide, N., & Takahashi, H. (2004). A decision based one-against-one method for multi-class support vector machine. *Pattern Analysis and Its Applications*, *7*(2), 164–175.

Dietterich, T. G., & Bakiri, G. (1995). Solving multiclass learning problems via error-correcting output codes. *The Journal of Artificial Intelligence Research*, *2*, 263–286.

Dreiseitl, S., Ohno-Machado, L., & Binder, M. (2000). Comparing three-class diagnostic tests by three-way ROC analysis. *Medical Decision Making*, *20*, 323–331.

Edwards, D., Metz, C., & Kupinski, M. (2005). The hypervolume under the ROC hypersurface of 'near-guessing' and 'near-perfect' observers in n-class classification tasks. *IEEE Transactions on Medical Imaging*, *24*(3), 293–299.

Fagin, R., Kumar, R., Mahdian, M., Sivakumar, D., & Vee, E. (2004). Comparing and aggregating rankings with ties. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems, PODS '04* (pp. 47–58).

Ferri, C., Hernández-Orallo, J., & Salido, M. (2003). Volume under the ROC surface for multi-class problems. In *Proceedings of 14th European conference on machine learning* (pp. 108–120).

Fieldsend, J., & Everson, R. (2005). Formulation and comparison of multi-class ROC surfaces. In *Proceedings of the ICML 2005 workshop on ROC analysis in machine learning* (pp. 41–48).

Fieldsend, J., & Everson, R. (2006). Multi-class ROC analysis from a multi-objective optimisation perspective. *Pattern Recognition Letters*, *27*, 918–927.

Flach, P. (2004). *Tutorial*: "*the many faces of ROC analysis in machine learning*". *Part III* (Technical report). International conference on machine learning 2004.

Frank, A., & Asuncion, A. (2010). UCI machine learning repository.

Freund, Y., Iyer, R. D., Schapire, R. E., & Singer, Y. (2003). An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, *4*, 933–969.
Fürnkranz, J. (2002). Round robin classification. *Journal of Machine Learning Research*, *2*, 721–747.
Fürnkranz, J., Hüllermeier, E., & Vanderlooy, S. (2009). Binary decomposition methods for multipartite ranking. In *Proceedings of the European conference on machine learning and knowledge discovery in databases: Part I, ECML PKDD '09* (pp. 359–374).
Hand, D., & Till, R. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, *45*(2), 171–186.
Hastie, T., & Tibshirani, R. (1998). Classification by pairwise coupling. *The Annals of Statistics*, *26*(2), 451–471.
Herbrich, R., Graepel, T., & Obermayer, K. (2000). Large margin rank boundaries for ordinal regression. In *Advances in large margin classifiers* (pp. 115–132). Cambridge: MIT Press.
Higgins, J. (2004). *Introduction to modern nonparametric statistics*. N. Scituate: Duxbury Press.
Hudry, O. (2008). NP-hardness results for the aggregation of linear orders into median orders. *Annals of Operations Research*, *163*, 63–88.
Huhn, J., & Hüllermeier, E. (2008). Is an ordinal class structure useful in classifier learning? *International Journal of Data Mining, Modelling and Management*, *1*(1), 45–67.
Kramer, S., Pfahringer, B., Widmer, G., & Groeve, M. D. (2001). Prediction of ordinal regression trees. *Fundamenta Informaticae*, *47*, 1001–1013.
Laguna, M., Marti, R., & Campos, V. (1999). Intensification and diversification with elite tabu search solutions for the linear ordering problem. *Computers and Operations Research*, *26*(12), 1217–1230.
Landgrebe, T., & Duin, R. (2006). A simplified extension of the area under the ROC to the multiclass domain. In *Seventeenth annual symposium of the pattern recognition association of South Africa* (pp. 241–245).
Lebanon, G., & Lafferty, J. (2002). Conditional models on the ranking poset. In *Advances in neural information processing systems* (Vol. 15, pp. 415–422).
Lehmann, E., & Romano, J. P. (2005). *Testing statistical hypotheses*. Berlin: Springer.
Li, J., & Zhou, X. (2009). Nonparametric and semiparametric estimation of the three way receiver operating characteristic surface. *Journal of Statistical Planning and Inference*, *139*, 4133–4142.
Mandhani, B., & Meila, M. (2009). Tractable search for learning exponential models of rankings. *Journal of Machine Learning Research. Proceedings Track*, *5*, 392–399.
Meila, M., Phadnis, K., Patterson, A., & Bilmes, J. (2007). Consensus ranking under the exponential model. In *Proceedings of the twenty-third conference annual conference on uncertainty in artificial intelligence (UAI-07)* (pp. 285–294).
Mossman, D. (1999). Three-way ROCs. *Medical Decision Making*, *19*(1), 78–89.
Nakas, C., & Yiannoutsos, C. (2004). Ordered multiple-class ROC analysis with continuous measurements. *Statistics in Medicine*, *23*(22), 3437–3449.
Pahikkala, T., Tsivtsivadze, E., Airola, A., Boberg, J., & Salakoski, T. (2007). Learning to rank with pairwise regularized least-squares. In *Proceedings of SIGIR 2007 workshop on learning to rank for information retrieval* (pp. 27–33).
Pepe, M. (2003). *Statistical evaluation of medical tests for classification and prediction*. Oxford: Oxford University Press.
Rajaram, S., & Agarwal, S. (2005). Generalization bounds for k-partite ranking. In *NIPS workshop on learning to rank*.
Robbiano, S. (2010). *Note on confidence regions for the ROC surface* (Technical report). Telecom ParisTech.
Rudin, C., Cortes, C., Mohri, M., & Schapire, R. E. (2005). Margin-based ranking and boosting meet in the middle. In *Proceedings of the 18th annual conference on learning theory, COLT'05* (pp. 63–78). Berlin: Springer.
Scurfield, B. (1996). Multiple-event forced-choice tasks in the theory of signal detectability. *Journal of Mathematical Psychology*, *40*, 253–269.
Tsybakov, A. (2004). Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, *32*(1), 135–166.
Vapnik, V. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, *10*(5), 988–999.
Venkatesan, G., & Amit, S. (1999). Multiclass learning, boosting, and error-correcting codes. In *Proceedings of the twelfth annual conference on computational learning theory, COLT'99* (pp. 145–155).
Waegeman, W., & Baets, B. D. (2011). On the era ranking representability of pairwise bipartite ranking functions. *Artificial Intelligence*, *175*, 1223–1250.
Waegeman, W., Baets, B. D., & Boullart, L. (2008a). On the scalability of ordered multi-class ROC analysis. *Computational Statistics and Data Analysis*, *52*, 3371–3388.
Waegeman, W., Baets, B. D., & Boullart, L. (2008b). ROC analysis in ordinal regression learning. *Pattern Recognition Letters*, *29*, 1–9.
Wakabayashi, Y. (1998). The complexity of computing medians of relations. *Resenhas*, *3*(3), 323–349.