Ranking Multi-Class Data: Optimality and Pairwise Aggregation

Stéphan Clémençon, Sylvain Robbiano, Nicolas Vayatis

Ranking Multi-Class Data: Optimality and Pairwise Aggregation

Stéphan Clémençon · Sylvain Robbiano · Nicolas Vayatis

Abstract It is the primary purpose of this paper to set the goals of ranking in a multiple-class context rigorously, following in the footsteps of recent results in the bipartite framework. Under specific *likelihood ratio monotonicity* conditions, optimal solutions for this global learning problem are described in the ordinal situation, *i.e.* when there exists a natural order on the set of labels. Criteria reflecting ranking performance under these conditions such as the ROC surface and its natural summary, the volume under the ROC surface (VUS), are next considered as targets for empirical optimization. Whereas *plug-in* techniques or the *Empirical Risk Maximization* principle can be then easily extended to the ordinal multi-class setting, reducing the K-partite ranking task to the solving of a collection of bipartite ranking problems, following in the footsteps of the *pairwise comparison* approach in classification, is in contrast more challenging. Here we consider a concept of *ranking rule consensus* based on the Kendall τ distance and show that, when it exists and is based on consistent ranking rules for the bipartite ranking subproblems defined by all consecutive pairs of labels, the latter forms a consistent ranking rule in the VUS sense under adequate conditions. This result paves the way for extending the use of recently developed learning algorithms, tailored for bipartite ranking, to multi-class data in a valid theoretical framework. Preliminary experimental results are presented for illustration purpose.

1 Introduction

Ranking is about learning to order observations so as to mimic the preorder induced by the (unknown) ordinal and discrete labels that are assigned to them, based on a set of labeled examples. This is an important issue in a wide variety of applications. In information retrieval for instance, the goal is to rank all possible documents by degree of relevance for a specific request, based on training data describing the characteristics X of a sample of documents and their relevance level through an ordinal discrete variable

Stéphan Clémençon - Telecom ParisTech - LTCI UMR Telecom ParisTech/CNRS No. 5141
E-mail: stephan.clemencon@telecom-paristech.fr
· Sylvain Robbiano - Telecom ParisTech - LTCI UMR Telecom ParisTech/CNRS No. 5141
· Nicolas Vayatis - ENS Cachan & UniverSud - CMLA UMR CNRS No. 8536

Y, that may take more than two values: in the LETOR benchmark data repository (see http://research.microsoft.com/en-us/um/people/letor/), it takes five values, ranging from 0 ("irrelevant) to 4 ("perfectly relevant"). In medicine as well, decision-making tools are also needed in a multiple-class setting, labels corresponding to an ordered gradation of illness (from "not diseased" to "seriously ill") and diagnostic test statistics are used for discriminating among disease states, see [Pep03], [DOMB00], [EMK05], [Mos99] or [NY04] for instance. This learning task, halfway between *classification* and *class distribution estimation*, presents a significant challenge to statisticians, precisely because of the nature of the object to guess, a preorder on the ensemble of possible observations. Obviously, there are many ways of comparing two preorders on a possibly continuous space and defining optimality criteria or risk measures in the ranking setup is not as straightforward as in *classification/regression* or in *density estimation*.

Whereas the issue of ranking data with binary labels, generally termed *bipartite ranking problem*, has recently been the subject of a good deal of attention in the statistical and machine-learning literature, leading to the design of novel efficient algorithms fully tailored for the ranking task (see [CV09e], [FISS03] and [CV09d] in particular) and giving rise to significant theoretical developments dedicated to this global learning problem (refer to [AGH$^+$05] or [CLV08] for instance), extension of related concepts and results to the ordinal multi-class context is far from immediate and poses many questions of theoretical or practical nature, not answered yet, see [Fla04] and the references therein. While, in the bipartite framework, the ROC curve (as well as transforms or summaries of the latter such as the celebrated AUC criterion) has provided the "definitive tool" for evaluating the relevance of ranking rules to a certain extent since its introduction in the 40's (*cf* [DS66]), it is only recently that this functional measure of accuracy has been generalized to the ordinal multi-class setup, leading to a specific notion of "ROC graph" tailored for K-partite ranking, see [Scu96]. Until now, the approach to ranking followed by most authors has consisted in optimizing a specific scalar criterion over a (nonparametric) set of ranking/scoring rules and applying the *empirical risk minimization* (ERM) paradigm. The "ranking risk" generally counts the number of "concordant pairs of observations" (*i.e.* the number of pairs of observations that are sorted in the same order as their labels) and takes the form of a U-statistic of degree two, see [CLV08], [RCMS05]. Alternately, in the bipartite framework, it may be a specific functional of the ranks induced by the ranking rule candidate, as in [Rud06], [CV07] or [CV09a].

The angle embraced in the present paper is quite different and our contribution to the analysis of the K-partite ranking problem is twofold. Its primary purpose is to describe the situation, in terms of data distribution, where a scoring rule that would be optimal for all bipartite ranking subproblems does exist, optimal scoring functions for the K-partite ranking problem being then naturally defined as those that define the same preorder on the input space as the latter. Here, we show that a *monotonicity likelihood ratio* assumption on the underlying collection of class distributions guarantees the existence of such optimal ranking rules. It is next shown that, under this assumption, the ROC graph or the volume it defines (generally termed the VUS criterion) can also be used for recovering the set of "optimal ranking rules" (originally defined without referring to these criteria) and, more generally, for quantifying ranking accuracy. In this respect, the K-partite framework with $K \geq 3$ is in contrast with the bipartite setup, where an optimal preorder on the input space always exists (*i.e.* that induced by the likelihood ratio of the two sole class distributions), which corresponds to a ROC

curve that dominates any other ROC curve in a pointwise manner. Once the goal of multi-class ranking has been set in a quantitative fashion, we turn to the secondary purpose of this article, namely the reduction of this learning problem to a series of bipartite ranking problems, with a modus operandi similar to the *pairwise comparison* method, also known as the "*all versus all*" approach (AVA), in multi-class pattern recognition, see [HT98] or [FÖ2] for instance. In continuity with the way optimality is defined here, the approach to K-partite ranking developed in this article lies in viewing it as a "superposition" of bipartite ranking tasks, following in the footsteps of the idea originally proposed in [FHV09]. Indeed, solutions are the ranking rules that are simultaneously optimal for the $K-1$ bipartite ranking problems related to all possible pairs of consecutive labels (*i.e.* pairs of consecutive class distributions). Hence, the ranking procedure we propose here is implemented in two stages. The first stage consists in solving the (bipartite) ranking subproblems separately, producing thus a collection of scoring rules. The second stage then involves the computation of a "median scoring rule", related to the collection obtained at the first stage and based on a specific notion of "distance" between scoring rules, disagreement being measured by the Kendall τ distance. It is shown that such a median always exists in the important situation where the scoring functions one seeks to summarize/aggregate are piecewise constant, and its computation is feasible. We next establish that the resulting consensus is a consistent ranking rule, provided that the ranking method used for solving the bipartite ranking subproblems is itself consistent. For completeness, alternative approaches, the *plug-in* method and techniques based on optimization of an empirical version of the VUS criterion namely, are also briefly mentioned. For simplicity, most results are stated in the case $K = 3$ only. Additionally, connections of K-partite ranking and ordinal regression, which involves a similar framework (*i.e.* also stipulates the existence of a natural order on the set of labels), are described. Although the material in the present paper is essentially theoretical, the principles investigated here are illustrated through a few numerical examples and several issues related to the practical implementation of the aggregation approach for multi-class ranking are discussed.

The rest of the paper is structured as follows. In section 2, the probabilistic setting is introduced, together with important notations, and the issue of ranking is formulated in an informal manner. A specific *monotonicity likelihood ratio* condition is stated, which is shown to guarantee the existence of a natural optimal preorder on the input space. In section 3, it is recalled how to extend the notion of ROC *graph* to the K-partite setup, with $K \geq 3$, and it is established that it provides a (functional) quantitative criterion that enables to assess the performance of any ranking rule candidate under the assumption aforementioned. It is also shown that the volume it defines in the ROC space, called the *volume under the* ROC *surface* (VUS) in the 3-class framework, may serve as a summary (scalar) criterion for ranking accuracy. Section 4 then describes possible approaches for multi-class ranking, the *plug-in* method and *empirical* VUS *maximization*, and highlights the fact that multi-class ranking can be viewed as a *multi-criteria optimization* task, whose each objective consists of a particular bipartite ranking subproblem. In section 5, a novel concept of *Kendall consensus* among ranking rules is introduced and it is proved that, when applied to a collection of $K-1$ ranking rules, where each of them is asymptotically optimal for a bipartite subproblem involving consecutive labels, it yields a consistent procedure in the VUS sense. Finally, section 6 displays some numerical results with the purpose to illustrate the principle of this aggregation approach. Technical proofs are deferred to the Appendix.

2 Theoretical Background - Preliminaries

We start off with a precise description of the probabilistic setup and an informal account of the ranking task in the ordinal multi-class context. We next detail a general framework where the goals of ranking can be rigorously set and ranking performance can be assessed in a quantitative manner, generalizing results established in the bipartite setting. Although K-partite ranking has already been tackled in previous works (see [QnL$^+$10], [FHV09] or [RA05]), from the perspective of (empirical) risk minimization mainly, to the best of our knowledge, no interpretable description of the set of optimal elements is available and no necessary and sufficient condition for the existence of optimal solutions (defined subsequently as simultaneous solutions of all bipartite ranking subproblems) has been stated explicitly yet, in the statistical and machine-learning literatures. It is the main purpose of this section to clarify these points.

2.1 Probabilistic setup - First notations

We place ourselves in the same probabilistic setup as that of *ordinal regression*. Precisely, one has a system consisting of a random output, taking its values in an ordered discrete set, $\mathcal{Y} = \{1, \ldots, K\}$ with $K \geq 2$ say, and a random input X, valued in a high-dimensional space $\mathcal{X}$, modelling some (hopefully relevant) information for predicting Y. Here and throughout, $F_k(dx)$ denotes X's conditional distribution given $Y = k$, $\mathcal{X}_k$ its support and we set $p_k = \mathbb{P}\{Y = k\}$ for $k = 1, \ldots, K$. With no restriction, we suppose that $\mathcal{X}$ coincides with $\cup_{k \leq K} \mathcal{X}_k$. Alternately, the distribution of the random pair (X, Y) can be described by X's marginal distribution $\mu(dx)$ and the posterior probabilities: $\eta_k(x) = \mathbb{P}\{Y = k \mid X = x\}$ with $x \in \mathcal{X}$ and $1 \leq k \leq K$ (notice that $\sum_{k=1}^{K} \eta_k \equiv 1$). For $k \in \{1, \ldots, K\}$, we also introduce the probability densities $\Phi_k(X) = dF_k/d\mu(X)$, as well as the (possibly infinite) likelihood ratios

$$\Phi_{k,l}(X) = \frac{dF_k}{dF_l}(X) = \frac{\Phi_k}{\Phi_l}(X),$$

with $1 \leq k,\ l \leq K$ and the convention that $u/0 = \infty$ for any $u \in]0, \infty[$ and $0/0 = 0$. These quantities are related to each other through the equations:

$$\mu(dx) = \sum_{k=1}^{K} p_k \cdot F_k(dx)$$

and, for $1 \leq k,\ l \leq K$,

$$\eta_k(X) = p_k \cdot \Phi_k(X) \text{ and } \Phi_{k,l}(X) = (p_l/p_k) \cdot \eta_k(X)/\eta_l(X).$$

The conditional expectation of the output random variable (we shall write "r.v." in abbreviated form) Y given X is denoted by:

$$\eta(X) = \mathbb{E}\left[Y \mid X\right] = \sum_{k=1}^{K} k \cdot \eta_k(X).$$

For any classifier $C : \mathcal{X} \to \{1, \ldots, K\}$, we set: $\forall (k, l) \in \{1, \ldots, K\}^2$,

$$\alpha_{k,l}(C) = \mathbb{P}\{C(X) = l \mid Y = k\}.$$

Here and throughout, $\mathcal{S}$ denotes the set of all borelian functions $s : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$. Its elements will be called *scoring functions*. Notice that the value $+\infty$ is allowed, so that likelihood ratios can be considered as scoring functions. The subset of real valued scoring functions is denoted by $\mathcal{S}_0$. Finally, Δ denotes the symmetric difference between sets, $\mathbb{I}\{\mathcal{E}\}$ the indicator function of any event $\mathcal{E}$ and $\mathcal{R}(\xi)$ the range of any mapping ξ.

2.2 Informal statement of the problem

In contrast to *multi-class pattern recognition* or *ordinal regression*, the goal we pursue here is not to predict the label Y attached to an observation X but to sort all instances $x \in \mathcal{X}$, by means of a *scoring function*[1] $s : \mathcal{X} \to]-\infty, +\infty]$ transporting the natural order on $\mathbb{R} \cup \{+\infty\}$ onto the space $\mathcal{X}$, in a way that

the random variables Y and $s(X)$ "tend to increase or decrease together".

Ideally, as the score $s(X)$ increases, with large probability, we would like to observe, as a majority, the instances with label $Y = 1$ first, those with label $Y = 2$ next, ...

In order to give a rigorous sense to this assertion, we introduce the following definition.

Definition 21. (Strict stochastic ordering) *Given two distribution functions $H(dt)$ and $G(dt)$ on $\mathbb{R} \cup \{+\infty\}$, it is said that $G(dt)$ is stochastically larger than $H(dt)$ iff for any $t \in \mathbb{R}$, we have $G(t) \leq H(t)$. We then write: $H \leq_{sto} G$. In addition, we will say that $G(dt)$ is strictly stochastically larger than $H(dt)$ iff it is stochastically larger and there exists $t \in \mathbb{R}$ such that $H(t) > G(t)$. In such a case, we will write: $H <_{sto} G$.*

Equipped with this notion, a minimal goal could naturally be to find a scoring function $s(x)$ so that the sequence of class distributions $(F_{s,k}(dt))_{1 \leq k \leq K}$ is *strictly stochastically increasing*:

$$F_{s,1} <_{sto} F_{s,2} <_{sto} \cdots <_{sto} F_{s,K}, \tag{1}$$

where $F_{s,k}(dt)$ denotes $s(X)$'s conditional distribution given $Y = k$ for $1 \leq k \leq K$. As shown by the following simplistic example, finding a scoring function $s(x)$ such that (1) holds is not always possible, even if the F_k's are all different.

Example 21. *Suppose that $\mathcal{X} = \{x_1, x_2, x_3\}$ and denote by $\delta_x(dx)$ the Dirac mass at x_k, $1 \leq k \leq 3$. Consider the following (pairwise distinct) distributions on $\mathcal{X}$: $\forall k \in \{1, 2, 3\}$*

$$F_k(dx) = \sum_{i=1}^{3} \omega_{k,i} \delta_{x_i}(dx),$$

where $\omega_{1,1} = \omega_{2,2} = \omega_{3,3} = 1/2$, $\omega_{1,2} = \omega_{2,3} = \omega_{3,1} = 1/3$ and $\omega_{1,3} = \omega_{2,1} = \omega_{3,2} = 1/6$. It is easy to check that, in this situation, the set of strict inequalities (1) is fulfilled for no scoring function s on $\mathcal{X}$. In contrast, the scoring function defined by $s(x_k) = k$ for $k \in \{1, 2, 3\}$ clearly fulfills the strict monotonicity property in the case where $\omega_{1,1} = \omega_{2,2} = \omega_{3,3} = 1/2$, $\omega_{1,2} = \omega_{2,1} = \omega_{3,2} = 1/3$ and $\omega_{1,3} = \omega_{2,3} = \omega_{3,1} = 1/6$ for instance.

[1] Any scoring function $s \in \mathcal{S}$ defines a preorder $\preceq_s$ on $\mathcal{X}$, defined by: $x \preceq_s x'$ iff $s(x) \leq s(x')$, for any $(x, x') \in \mathcal{X}^2$.

The example above shows that, in absence of any distributional assumption, it may happen no satisfactory solution to the K-partite ranking issue exists, *i.e.* it is not possible to define a preorder on $\mathcal{X}$ that permits to predict well that defined by the (unobserved) labels. In this respect, the K-partite ranking problem contrasts sharply with multi-class classification, that does not stipulate the existence of any predefined order on the set of output values and can be formulated in a universal manner, *i.e.* whatever the distribution of the pair (X, Y), there exists an optimal partition of the feature space. It is the purpose of the following subsection to define precisely the set of optimal scoring functions in the K-partite setup and to formulate a necessary and sufficient condition for this set to be non empty.

2.3 Assumptions and Optimality

The fact that the monotonicity property (1) may be fulfilled by no scoring function in certain situations as soon as $K \geq 3$ is in contrast with the so-termed *bipartite setup*, corresponding to the case $K = 2$ namely, where we always have

$$F_{s,1}(dt) <_{sto} F_{s,2}(dt),$$

when taking $s(x)$ as the likelihood ratio $dF_2/dF_1(x)$ for instance, as soon as $F_1(dx) \neq F_2(dx)$. In this situation, the notion of ROC curve and the related concept of AUC criterion (see [HM82]) enable to define a class of *optimal scoring functions*, which turns out to be the set of elements $s \in \mathcal{S}$ such that: $\forall (x, x') \in \mathcal{X}^2$, $dF_2/dF_1(x) < dF_2/dF_1(x') \Rightarrow s(x) < s(x')$. For clarity, we recall the following definition.

Definition 22. (ROC CURVE) *Let $F_1(dx)$ and $F_2(dx)$ be two probability distributions on $\mathcal{X}$. The* ROC *curve of a scoring function $s : \mathcal{X} \to]-\infty, +\infty]$ with respect to the pair (F_1, F_2) is the parametrized curve*

$$t \in \mathbb{R} \mapsto \left(\mathbb{P}\left\{s(X) > t \mid Y = 1\right\}, \mathbb{P}\left\{s(X) > t \mid Y = 2\right\}\right).$$

By convention, possible jumps (corresponding to points where the distributions $F_{s,1}(dt)$ and/or $F_{s,2}(dt)$ are degenerate) are connected by line segments, in order to guarantee the continuity of the curve. Equipped with this convention, the ROC curve may be viewed as the graph of a certain non decreasing *càd-làg*[2] (càd-làg standing for "continue à droite et limitée à gauche", *i.e.* right-continuous and left-limited) mapping $\alpha \in [0, 1] \mapsto \mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_2}(\mathrm{s}, \alpha)$, defined by

$$\mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_2}(\mathrm{s}, \alpha) = 1 - \mathrm{F}_{\mathrm{s},2} \circ \mathrm{F}_{\mathrm{s},1}^{-1}(1 - \alpha)$$

at points α such that $F_{s,1} \circ F_{s,1}^{-1}(1 - \alpha) = 1 - \alpha$, denoting by $W^{-1}(u) = \inf\{t \in] - \infty, +\infty] : W(t) \geq u\}$, $u \in [0, 1]$, the generalized inverse of any cdf $W(t)$ on $\mathbb{R} \cup \{+\infty\}$. Observe that it connects the point $(0, F_2(\mathcal{X}_2 \setminus \mathcal{X}_1))$ to $(F_1(\mathcal{X}_1 \setminus \mathcal{X}_2), 1)$ and that, in absence of plateau, the curve $\alpha \mapsto \mathrm{ROC}_{\mathrm{F}_2,\mathrm{F}_1}(\alpha)$ is the image of $\alpha \mapsto \mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_2}(\alpha)$ by the reflection with the line of Eq. "$\beta = \alpha$" as axis. We refer to Appendix A in [CV09e] for a list of properties of ROC curves (see Proposition 17 therein).

[2] Recall that, by definition, a *càd-làg* function $h : [0, 1] \to \mathbb{R}$ is such that $\lim_{s \to t,\ s<t} h(s) = h(t-) < \infty$ for all $t \in]0, 1]$ and $\lim_{s \to t,\ s>t} h(s) = h(t)$ for all $t \in [0, 1[$. Its completed graph is obtained by connecting the points $(t, h(t-))$ and $(t, h(t))$, when they are not equal, by a vertical line segment and thus forms a continuous curve.

Equipped with this concept, notice that $F_{s,1} \leq_{sto} F_{s,2}$ means that the curve is above the first diagonal of the unit square and it coincides with the latter when $F_1 = F_2$, whatever $s(x)$. More generally, the closer to the upper left corner of $[0,1]^2$ the curve $\mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_2}(\mathrm{s},.)$, the "stochastically larger" than $F_{s,1}$ the distribution $F_{s,2}$ can be viewed. Hence, ROC analysis induces a partial order on the set of scoring functions: with respect to the pair (F_1, F_2), a scoring functions $s(x)$ is said to be less accurate than another one $s'(x)$ when: $\forall \alpha \in [0,1]$,

$$\mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_2}(\mathrm{s},\alpha) \leq \mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_2}(\mathrm{s}',\alpha).$$

In regards to this way of evaluating ranking performance in the bipartite situation, the set $\mathcal{S}^*_{F_1,F_2}$ of optimal scoring functions is the set of functions $s \in \mathcal{S}$ such that:

$$\forall (x,x') \in \mathcal{X}^2 : \ \Phi_{F_2,F_1}(x) < \Phi_{F_2,F_1}(x') \Rightarrow s(x) < s(x').$$

This may be established by standard Neyman-Pearson arguments, we refer to Proposition 4 in [CV09e] for further details. One key advantage of ROC analysis in the context of binary classification lies in the fact that it permits to visualize the two types of error of a classifier $C(X)$ in a way that is insensitive to class skew, see [LF03]: the ROC curve of $C(X)$, viewed as a scoring function, is indeed the broken line that connect the points $(0,0)$, $(\alpha_{1,2}(C), 1-\alpha_{2,1}(C))$ and $(1,1)$.

We also recall that computing the *area under the* ROC *curve*, the quantity $\mathrm{AUC}_{\mathrm{F}_1,\mathrm{F}_2}(\mathrm{s}) = \int_{\alpha\in[0,1]} \mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_2}(\mathrm{s},\alpha)\mathrm{d}\alpha$ namely, is a widely used way of summarizing s's ranking performance with respect to (F_1, F_2). Beyond the fact that $\mathcal{S}^*_{F_1,F_2}$ naturally coincides with the set of scoring functions s with maximum AUC, the popularity of this criterion arises from its probabilistic interpretation as the (theoretical) "rate of concordant pairs" (in this respect, its empirical counterpart coincides with the two-sample Wilcoxon Mann-Whitney statistic). Indeed, we have:

$$\mathrm{AUC}_{\mathrm{F}_1,\mathrm{F}_2}(\mathrm{s}) = \mathbb{P}\left\{\mathrm{s}(\mathrm{X}) < \mathrm{s}(\mathrm{X}')\right\} + \frac{1}{2}\mathbb{P}\left\{\mathrm{s}(\mathrm{X}) = \mathrm{s}(\mathrm{X}')\right\},$$

where (X, X') denotes a pair of independent r.v.'s with respective marginal distributions $F_1(dx)$ and $F_2(dx)$.

Going back to the multiple-class setting, in accordance with the goal of K-partite ranking described above in an informal manner (see subsection 2.2), optimal scoring functions should be naturally defined as those belonging to the set

$$\mathcal{S}^* \stackrel{def}{=} \bigcap_{1\leq k<l\leq K} \mathcal{S}^*_{k,l},$$

where we set $\mathcal{S}^*_{k,l} = \mathcal{S}^*_{F_k,F_l}$ for notational convenience (note that $\mathcal{S}^* = \bigcap_{1\leq k<K} \mathcal{S}^*_{k,k+1}$).

Hence, by definition, optimal scoring for the K-partite ranking problem are those that are simultaneously optimal for the $K(K-1)/2$ bipartite ranking subproblems defined by all possible pairs of distinct label values. Given the definition of the set of optimal elements, anticipating the second part of the paper (see also [FHV09]), a natural approach to K-partite ranking in practice could be implemented in two stages: solve first the bipartite subproblems independently and next try to find a scoring function that induces a "barycentric" preorder on $\mathcal{X}$, as close as possible (in a sense

that will be specified later) to those induced by the bipartite solutions produced at the first stage.

As highlighted by Example 21, the set $\mathcal{S}^*$ may be empty. The following assumption, stipulating that all ratios $\Phi_{k,l}$, $1 \leq l < k \leq K$, increase or decrease together, can be easily seen to be a necessary and sufficient condition for $\mathcal{S}^*$ to be non empty. Notice incidentally that, since one may write $\Phi_{k,l} = \prod_{i=l}^{k-1} \Phi_{i,i+1}$ for all $1 \leq l < k \leq K$, assuming that the $\Phi_{k,k+1}$'s all vary in the same direction guarantees that this is also the case for the collection of ratios $\Phi_{k,l}$, $1 \leq l < k \leq K$.

Assumption 1. *For any $(k,l) \in \{1, \ldots, K-1\}^2$, all $(x,x') \in \mathcal{X}^2$, we have:*

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

Indeed, we have the following result, for which the proof is straightforward and thus omitted.

Theorem 21. *The set $\mathcal{S}^*$ is non empty if and only if Assumption 1 is fulfilled. In such a case, we necessarily have*

$$\mathcal{X}_{k'} \cap \mathcal{X}_{l'} \subset \mathcal{X}_k \cap \mathcal{X}_l \text{ for any } 1 \leq k' \leq k < l \leq l' \leq K.$$

In addition, the set $\mathcal{S}^*$ is the set of scoring functions $s \in \mathcal{S}$ such that,

$$\forall (x,x') \in \mathcal{X}^2 : \ \exists k \in \{1, \ldots, K-1\} \text{ such that } \Phi_{k+1,k}(x) < \Phi_{k+1,k}(x') \Rightarrow s(x) < s(x').$$

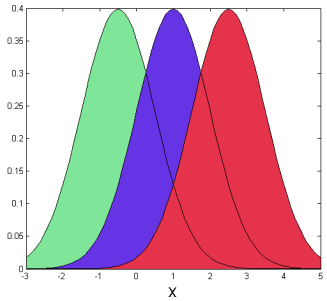
We point out that a related condition, called *ERA ranking representability*, has been introduced in [WB11], see Definition 2.1 therein. Precisely, it can be easily checked that Assumption 1 means that the collection of (bipartite) ranking functions $\{\Phi_{k+1,k} : 1 \leq k < K\}$ is an ERA ranking representable set of ranking functions.

In order to gain insight into the meaning of Assumption 1, it may be useful to exhibit (simple but illustrative) examples for which it is (not) satisfied (see also the toy example given in Section 6).

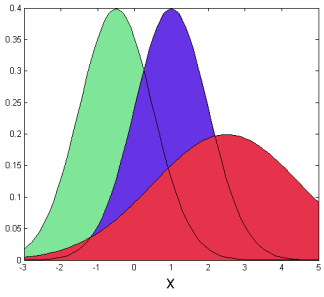
Example 22. *Consider the situation where the output variable Y can take $K = 3$ values: 1 ("bad"), 2 ("average") and 3 ("good") and ranking must be based on observations of a random variable X, such that its conditional distribution given $Y = k$, $1 \leq k \leq 3$, is the Gaussian $\mathcal{N}(m_k, \sigma_k^2)$ with mean $m_k \in \mathbb{R}$ and variance $\sigma_k^2 > 0$. As illustrated by Fig. 1 a below, when $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$, it is straightforward that Assumption 1 is fulfilled iff we have either $m_1 \leq m_2 \leq m_3$ or else $m_3 \leq m_2 \leq m_1$. Fig. 1 b depicts a situation where $m_1 < m_2 < m_3$ and $\sigma_3^2 > \sigma_2^2 = \sigma_1^2$ and for which the random observation X does not permit to recover the preorder induced by the output variable in a satisfactory manner (with the parameter values indicated, one may easily check that the condition involved in Assumption 1 is not satisfied for $(x,x') = (-2,1)$ for instance).*

Before tackling questions related to the design of quantitative performance criteria for the multiple-class ranking problem with $\mathcal{S}^*$ as set of optimal elements, a few remarks are in order.

Remark 21. (Separable case) *We point out that in the case where the supports $\mathcal{X}_k$, $1 \leq k \leq K$, are pairwise disjoints up to μ-negligible sets, Assumption 1 is clearly fulfilled (for $k \neq l$, $\Phi_{l,k}(X)$ is then equal either to 0 or else to ∞). Obviously, in this case, any solution to the multi-class classification problem, i.e. any classifier $C(X)$ such that $\mathbb{P}\{Y \neq C(X)\} = 0$, is also an optimal scoring function.*



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

Remark 22. (Multi-class ranking as bipartite ranking) *Notice that, if all the class distributions $F_1(dx)$, ..., $F_K(dx)$ have same support and a stronger version of Assumption 1 is fulfilled, stipulating that the strict ordering is actually preserved, $\mathcal{S}^*$ coincides with all the $\mathcal{S}^*_{k,l}$'s, $k < l$. In that case, and in that case solely, performing the multi-class ranking task theoretically boils down to solving a single bipartite ranking problem, related to any pair of class distributions (F_k, F_l), $k < l$. From a practical perspective however, even in such a case, using all the data in the learning stage, as in the procedure proposed in section 5, generally increases the accuracy obtained in practice, see section 6 for an illustrative example.*

The next result reveals that Assumption 1 boils down to supposing that the family of densities $\{\Phi_k(x) : 1 \leq k \leq K\}$ has *monotone likelihood ratio*. Its proof is obvious and left to the reader.

Proposition 22. *Assumption 1 is fulfilled if and only if there exists a real valued borelian function $s^*(x)$ such that for any $k < l$ in $\{1, \ldots, K\}$, the ratio $\Phi_{l,k}(x)$ is a non decreasing function of $s^*(x)$. In this case, the scoring function $s^*(x)$ belongs to the set $\mathcal{S}^*$.*

2.4 Connection 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* [Aga08]. In this case, it is shown that the optimal classifier can be built by thresholding the regression function at specific levels $t_0 = 0 < t_1^* < \ldots < t_{K-1}^* < 1 = t_K$, 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 $t_0 = 0 < t_1 < \ldots < t_{K-1} < 1 = t_K$ 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 [Aga08]. One may refer to [KPWG01] in the case of regression trees for instance.
The next theorem shows that, when Assumption 1 is fulfilled, the regression function $\eta(x)$ lies in the set $\mathcal{S}^*$, making ranking methods based on the use of the latter as a scoring function legitimate (see section 3 below for further details on the *plug-in* approach). Notice however that thresholding ranking functions is not the sole possible approach to ordinal regression, see [TDFMSN06] or [HH09] for instance.

Theorem 23. *If Assumption 1 is satisfied, the regression function $\eta(x)$ then belongs to the set of optimal scoring functions $\mathcal{S}^*$.*

This novel result shows in particular that, under Assumption 1, the optimal prediction rule $C^*(x)$ can be obtained by thresholding adequately any scoring function in $\mathcal{S}^*$. Hence, in the approach to ordinal regression described above, the regression estimation stage could be replaced by a ranking procedure, producing a nearly optimal scoring function. Incidentally, we point out that this result legitimates the use of ranking criteria such as the ROC graph or summaries of the latter in the context of ordinal regression, as proposed in [WBB08b] for instance.

3 Assessing performance in K-partite ranking

Now that the situation where multi-class ranking can be considered as a well-posed problem (*i.e.* where optimal ranking rules do exist) has been made explicit, we show here how ranking performance can be quantitatively assessed in the K-partite setting, with $K \geq 3$. Precisely, we prove that, under Assumption 1, the set of optimal ranking rules $\mathcal{S}^*$ coincides with the set of optima of a functional criterion, referred to as the ROC *graph* and which extends the notion of ROC curve (see Definition 22) to the K-partite situation. Though desirable and expected both at the same time, such a result is crucial, insofar as the optimal set $\mathcal{S}^*$ has been defined without referring to any quantitative criterion in section 2.

3.1 Multiple-class ROC analysis

In the bipartite setup, ROC analysis is the most standard, and somehow ultimate, way of evaluating ranking performance, see [Faw06]. Alternatives consist either of different parametrizations, such as the *Precision-Recall* curve, which offers a scale-adapted graphical display that allows for visualizing ranking performance more easily in case of a highly skewed pooled distribution (refer to [CV09b] for instance), or else of summaries of the ROC curve: AUC, local AUC, p-norm push, *etc.* See [CV09a] or [CV07] and the references therein.

Following in the footsteps of [Scu96] in situations with more than two classes, the ROC graphic of a scoring function $s(x)$ becomes the set of points

$$M(\mathbf{t}) = \left(F_{s,1}(t_1) - F_{s,1}(t_0),\ \ldots,\ ,\ F_{s,K}(t_K) - F_{s,K}(t_{K-1})\right), \tag{2}$$

where $-\infty = t_0 < t_1 \leq \ldots \leq t_{K-1} < t_K = \infty$, incidentally we will denote by $\mathcal{T}_K$ the set of all such vectors $\mathbf{t} = (t_1,\ \ldots,\ t_{K-1})$ in the following. Notice that $F_{s,K}(t_K) = 1$ and $F_{s,1}(t_0) = 0$ and that the coordinates of the point (2) coincides with the diagonal entries of the *confusion matrix* of the classification rule defined by thresholding $s(x)$ at the levels t_k, $1 \leq k < K$,

$$C_{s,\mathbf{t}}(x) = \sum_{k=1}^{K} k \cdot \mathbb{I}\{t_{k-1} < s(x) \leq t_k\}.$$

We have indeed $\mathbb{P}\left\{C_{s,\mathbf{t}}(X) = k \mid Y = k\right\} = F_{s,k}(t_k) - F_{s,k}(t_{k-1})$ for all k in $\{1,\ \ldots,\ K\}$.

Remark 31. (On graph conventions.) *It should be pointed up that, in the case $K = 2$, the* ROC *graphic defined above does not coincide with the* ROC *curve defined in subsection 2.3 (see Definition 22) but with its image by the transform $(\alpha, \beta) \in [0,1]^2 \mapsto (1 - \alpha, \beta)$.*

As in the bipartite situation, we connect by convention all possible discontinuities (due to possible jumps of the distributions $F_{s,k}$) by parts of affine hyperplanes, so that the ROC graphic is a continuous manifold of dimension $K - 1$. We thus call it "ROC manifold". In order to lighten notation, we take $K = 3$ in the following. We then call ROC space the unit cube $[0,1]^3$ equipped with the usual $\alpha\beta\gamma$ cartesian coordinate system.

The ROC surface clearly concatenates all the information carried by the three curves $\mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_2}(\mathrm{s},.)$, $\mathrm{ROC}_{\mathrm{F}_2,\mathrm{F}_3}(\mathrm{s},.)$ and $\mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_3}(\mathrm{s},.)$. In particular, the intersection of the surface with each of the three facets of the positive orthant coincides with the image of one of these curves by a simple transform, see Proposition 31 below. The ROC graph is then a *parametric surface*, that coincides with the (completed) graph

$$\left\{(\alpha, \mathrm{ROC}(\mathrm{s},\alpha,\gamma), \gamma) :\ (\alpha,\gamma) \in [0,1]^2 \text{ such that } \gamma \leq \mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_3}(\mathrm{s}, 1-\alpha)\right\}$$

of a mapping $\mathrm{ROC}(\mathrm{s},.,.)$ defined on the set

$$\mathcal{I}_s \stackrel{def}{=} \left\{(\alpha,\gamma) \in [0,1]^2 :\ \gamma \leq \mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_3}(\mathrm{s}, 1-\alpha)\right\}$$

and such that, at any point (α, γ) for which $\gamma \leq \mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_3}(\mathrm{s}, 1-\alpha)$, $F_{s,1} \circ F_{1,s}^{-1}(\alpha) = \alpha$ and $F_{s,3} \circ F_{3,s}^{-1}(\gamma) = \gamma$, we have:

$$\mathrm{ROC}(\mathrm{s},\alpha,\gamma) = \mathrm{F}_{2,\mathrm{s}} \circ \mathrm{F}_{3,\mathrm{s}}^{-1}(1-\gamma) - \mathrm{F}_{2,\mathrm{s}} \circ \mathrm{F}_{1,\mathrm{s}}^{-1}(\alpha). \tag{3}$$

By convention, we extend the mapping $\mathrm{ROC}(\mathrm{s},.,.)$ on the whole unit square $[0,1]^2$ by setting $\mathrm{ROC}(\mathrm{s},\alpha,\gamma) = 0$ for any (α,γ) such that $\gamma > \mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_3}(\mathrm{s}, 1-\alpha)$. Hence we may write:

$$\forall (\alpha,\gamma) \in [0,1]^2,\ \mathrm{ROC}(\mathrm{s},\alpha,\gamma) = \left(\mathrm{F}_{2,\mathrm{s}} \circ \mathrm{F}_{3,\mathrm{s}}^{-1}(1-\gamma) - \mathrm{F}_{2,\mathrm{s}} \circ \mathrm{F}_{1,\mathrm{s}}^{-1}(\alpha)\right)_+,$$

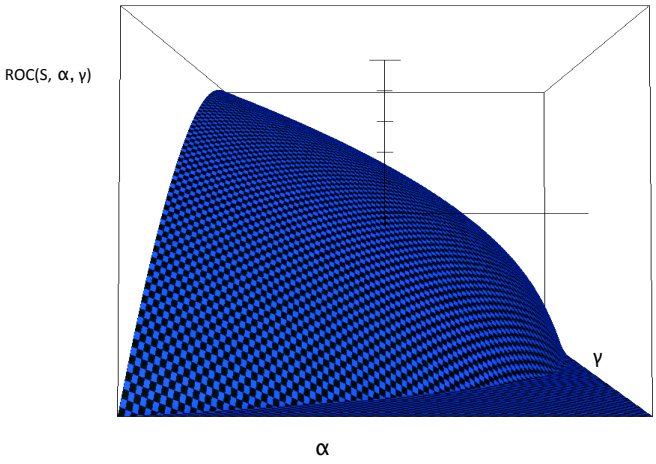


Fig. 2 Plot of the ROC surface of a scoring function s: $(\alpha,\ \gamma) \in (0,1)^2 \mapsto \mathrm{ROC}(\mathrm{s}, \alpha, \gamma)$

where $u_+ = \max(0,\ u)$ denotes the positive part of any real number u. The advantage of this representation is obvious when it comes to comparing the ROC graphs of two different scoring functions $s(x)$ and $s'(x)$ (since $\mathcal{I}_s \neq \mathcal{I}_{s'}$ in general).

For notational simplicity's sake, we omit to index $\mathrm{ROC}(\mathrm{s}, ., .)$ by the distributions $F_1(dx)$, $F_2(dx)$ and $F_3(dx)$ on which it depends throughout the paper.

We point out that different notions of ROC graph 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 [FHOS03] (see also [FE05], [FE06] and [HT01]) from a *one-versus-one* angle or in [Fla04] 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 [WBB08b].

Remark 32. (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 $H_k(dx)$ and $G_k(dx)$, for $k \in \{1,\ \ldots,\ K-1\}$. This boils down to plot the graph of the mapping $\alpha \in [0,1] \mapsto (\mathrm{ROC}_{\mathrm{H}_1,\mathrm{G}_1}(\mathrm{s},\alpha),\ \ldots,\ \mathrm{ROC}_{\mathrm{H}_{K-1},\mathrm{G}_{K-1}}(\mathrm{s},\alpha))$. 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: $\mathrm{ROC}_{\mathrm{H}_k,\mathrm{G}_k}(\mathrm{s},\alpha) \leq \mathrm{ROC}_{\mathrm{H}_k,\mathrm{G}_k}(\mathrm{s}^*,\alpha)$ for all $\alpha \in [0,1]$, $1 \leq k < K$. The likelihood ratio $dG_k/dH_k(X)$ is indeed a non decreasing function of $s^*(X)$, see Theorem 3.4.1 in [LR05] 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 $H_k(dx)$ and*

$G_k(dx)$ depend on the latter, they can be expressed as $\sum_{l\leq k} p_l F_l(dx)/(\sum_{l\leq k} p_l)$ and $\sum_{l>k} p_l F_l(dx)/(\sum_{l>k} p_l)$ respectively.

Remark 33. (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 $(1, 0, 0)$, writing abusively $\alpha_{k,l}$ for $\alpha_{k,l}(C)$ here, for notational simplicity. 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 [FHOS03]. 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.*

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 [CV09e] and the definition of the ROC surface given in Eq. (3), it is thus left to the reader.

Proposition 31. (Properties of the ROC surface) *For any distributions $F_1(dx)$, $F_2(dx)$ and $F_3(dx)$ 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 $\{(\alpha, \mathrm{ROC}(\mathrm{s}, \alpha), \gamma)\}$ with the plane of Eq. "$\alpha = 0$" coincides with the curve $\{(\beta, \mathrm{ROC}_{\mathrm{F}_2,\mathrm{F}_3}(\mathrm{s}, \beta))\}$ 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 $\{(\alpha, \mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_3}(\mathrm{s}, \alpha))\}$ by the mapping $\psi(\alpha, \gamma)$ and that with the plane of Eq. "$\gamma = 0$" to the image of $\{(\alpha, \mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_2}(\mathrm{s}, \alpha))\}$ 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 $(\alpha, \gamma) \in [0, 1]^2$:*

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

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 $\mathrm{ROC}^* \stackrel{\mathrm{def}}{=} \mathrm{ROC}(\mathrm{s}^*, ., .)$, 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 $dF_2/dF_1(x)$ and $dF_3/dF_2(x)$ are constant on the subsets $\{x \in \mathcal{X} / \ Q(dF_2/dF_1(X), \alpha_2) \leq dF_2/dF_1(x) \leq Q(dF_2/dF_1(X), \alpha_1)\}$ and $\{x \in \mathcal{X} / \ Q(dF_3/dF_2(X), \gamma_2) \leq dF_3/dF_2(x) \leq Q(dF_3/dF_2(X), \gamma_1)\}$ respectively, denoting by $Q(Z, \alpha)$ the quantile of order $1 - \alpha$ of any random variable Z.*

5. **Differentiability.** *Assume that the distributions $F_1(dx)$, $F_2(dx)$ and $F_3(dx)$ 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: $\forall(\alpha,\gamma) \in \mathcal{I}_s$,*

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

Preliminary results related to statistical estimation of the ROC surface of a fixed scoring function $s(x)$ can be found in [LZ09], additional results related to the building of confidence regions in the ROC space $[0,1]^3$ are established in [Rob10].

A partial preorder on $\mathcal{S}$. The ROC surface provides a visual tool for comparing the ranking performance of two scoring functions: we shall say that a scoring function $s(x)$ provides a better ranking than $s'(x)$ when: $\forall(\alpha,\gamma) \in [0,1]^2$,

$$\mathrm{ROC}(\mathrm{s},\alpha,\gamma) \geq \mathrm{ROC}(\mathrm{s}',\alpha,\gamma).$$

This criterion induces a partial order over the space of all scoring functions $\mathcal{S}$, for which $\mathcal{S}^* = \mathcal{S}_{1,2}^* \cap \mathcal{S}_{2,3}^*$ appear as the set of optimal elements. We highlight the fact that this is non trivial since the set $\mathcal{S}^*$ has been defined with no reference to the concept of 3-partite ROC graph. To see this, it suffices to observe that the ROC surface of $s(x)$ can be expressed in terms of ROC curves, as follows: $\forall(\alpha,\gamma) \in [0,1]^2$,

$$\mathrm{ROC}(\mathrm{s},\alpha,\gamma) = \left(\mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_2}(\mathrm{s},1-\alpha) - \mathrm{ROC}_{\mathrm{F}_3,\mathrm{F}_2}(\mathrm{s},\gamma)\right)_+.$$

Hence, optimizing the ROC surface consists of simultaneously optimizing the ROC curves related to the two pairs of distributions (F_1,F_2) and (F_2,F_3). The next theorem immediately results from this observation, it states that the ROC surface provides a (functional) quantitative means for assessing ranking accuracy.

Theorem 32. *Suppose that Assumption 1 is fulfilled and set* $\mathrm{ROC}^*(.,.) = \mathrm{ROC}(\mathrm{s}^*,.,.)$ *for $s^* \in \mathcal{S}^*$. We have, for any scoring function $s \in \mathcal{S}$ and for all $(\alpha,\gamma) \in [0,1]^2$,*

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

In addition, if we set, for any $\alpha \in [0,1]$, $k \in \{1,\ 2,\ 3\}$ and $s \in \mathcal{S}$,

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

where $Q^{(i)}(s,\alpha)$ denotes the quantile of order α of $s(X)$'s conditional distribution given $Y = i$ and assume that $\eta(X)$ is a continuous random variable, we have: $\forall(\alpha,\gamma) \in [0,1]^2$,

$$\mathrm{ROC}^*(\alpha,\gamma) - \mathrm{ROC}(\mathrm{s},\alpha,\gamma) \leq \mathbb{I}\{(\alpha,\gamma) :\ \gamma \leq \mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_3}^*(1-\alpha)\} \cdot \left(\Theta_1(\mathrm{s},\alpha) + \Theta_2(\mathrm{s},\gamma)\right),$$

where

$$\Theta_1(s,\alpha) = \frac{\mathbb{I}\{\alpha \neq 0\}}{p_2 Q^{(1)}(\eta_1,\alpha)}\mathbb{E}\left[\left|\eta_1(x) - Q^{(1)}(\eta_1,\alpha)\right| \cdot \mathbb{I}\{R_{s^*,\alpha}^{(1)} \Delta R_{s,\alpha}^{(1)}\}\right],$$
$$\Theta_2(s,\gamma) = \frac{\mathbb{I}\{\gamma \neq 1\}}{p_2 Q^{(3)}(\eta_3,1-\gamma)}\mathbb{E}\left[\left|\eta_3(X) - Q^{(3)}(\eta_3,1-\gamma)\right| \cdot \mathbb{I}\{R_{s^*,1-\gamma}^{(3)} \Delta R_{s,1-\gamma}^{(3)}\}\right],$$

for any $s^* \in \mathcal{S}^*$.

In the case where $s(x)$ 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 $\mathrm{ROC}(\mathrm{s},\alpha,\gamma) = 1-\alpha-\gamma$. By contrast, in the separable situation (see Remark 21), the optimal ROC surface coincides with the surface of the unit cube $[0,1]^3$. Though stated in a restrictive setting (stipulating the continuity of $\eta(X)$'s distribution) for simplicity's sake, the second part of the theorem above reveals that the pointwise difference between the optimal ROC surface and that of a candidate s is related to the errors made in recovering the level sets $R^{(1)}_{s^*,\alpha}$ and $R^{(3)}_{s^*,1-\gamma}$ through $R^{(1)}_{s,\alpha}$ and $R^{(3)}_{s,1-\gamma}$.

The following result establishes that, reciprocally, if there exists some scoring function whose ROC surface dominates in a pointwise manner any other ROC surface, then it belongs to $\mathcal{S}^*$ and Assumption 1 is thus necessarily fulfilled.

Theorem 33. *Suppose that there exists $s^* \in \mathcal{S}$ such that, for any scoring function $s \in \mathcal{S}$, we have: $\forall(\alpha,\gamma) \in [0,1]^2$,*

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

Then, the set $\mathcal{S}^* = \mathcal{S}^*_{1,2} \cap \mathcal{S}^*_{2,3}$ is non empty and the scoring function s^* belongs to it.

Though simple, the results stated in this subsection are crucial, since they provide theoretical grounds for the use of ROC analysis in the context of K-partite ranking. In summary, they show that the scoring functions that are optimal for all bipartite ranking subproblems (when they exist, that is to say iff Assumption 1 is satisfied, *cf* section 2) are the optimal elements for the ROC surface criterion, and that, reciprocally, if the ROC surface criterion has an optimum (*i.e.* there exists a scoring function whose ROC surface dominates any other ROC surface in a pointwise fashion), then Assumption 1 is necessary fulfilled (and the optimum aforementioned thus belongs to the set $\mathcal{S}^*$).

3.2 Volume under the ROC surface: a summary of ranking performance

In a manner similar to the bipartite situation, where the AUC criterion provides a popular scalar summary of the ROC curve, one may consider the *volume under the* ROC *surface* (VUS in abbreviated form) in the three-class framework, see [Scu96]. As stated in the following proposition, this induces a total preorder on the set of scoring functions, for which $\mathcal{S}^*$ correspond to the set of optimal elements. One may refer to [LD06] and [FHOS03] for arguments in favor of the use of this criterion in the classification context too, and to [WBB08b] in the ordinal regression setup. As mentioned in subsection 3.1, other notions notions of ROC graph can be found in the literature, leading to other summary quantities, also referred to as VUS, such as that introduced in [HT01].

Proposition 34. (VUS CRITERION) *Let $s(x)$ be a scoring function. The volume under its* ROC *surface is:*

$$\begin{aligned}
\mathrm{VUS}(\mathrm{s}) &= \int\int \mathrm{ROC}(\mathrm{s},\alpha,\gamma)\mathrm{d}\alpha\mathrm{d}\gamma, \\
&= \int_{\alpha=0}^{1} \mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_2}(\mathrm{s},1-\alpha)\mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_3}(\mathrm{s},1-\alpha)\mathrm{d}\alpha \\
&\quad - \int_{\gamma=0}^{1} \mathrm{ROC}_{\mathrm{F}_3,\mathrm{F}_2}(\mathrm{s},\gamma)\left(1-\mathrm{ROC}_{\mathrm{F}_3,\mathrm{F}_1}(\mathrm{s},\gamma)\right)\mathrm{d}\gamma.
\end{aligned}$$

Under Assumption 1, we have: $\forall s \in \mathcal{S}$,

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

with $\mathrm{VUS}^* = \mathrm{VUS}(\mathrm{s}^*)$ *for* $s^* \in \mathcal{S}^*$.

This proposition is actually a corollary of Theorem 32, its proof is thus omitted. Notice in particular that, when $F_{s,1} = F_{s,2} = F_{s,3}$, that is to say when the scoring function s has no capacity to discriminate between the three classes, we have VUS(s) = 1/6, the ROC surface then corresponds to the area delineated by the triangle with vertices $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$. In contrast to this situation, in the case where the $F_{s,k}$'s have pairwise disjoint supports, the VUS can reach the value 1 (respectively, the value 0), when, in addition, for all $(x_1, x_2, x_3) \in \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3$, we have $s(x_1) < s(x_2) < s(x_3)$ (respectively, $s(x_3) < s(x_2) < s(x_1)$).

Like the AUC criterion, VUS(s) can be interpreted in a probabilistic manner. For clarity, we recall the following result.

Proposition 35. ([Scu96]) *For any scoring function* $s \in \mathcal{S}$, *we have:*

$$\begin{aligned}\mathrm{VUS(s)} = \mathbb{P}\left\{\mathrm{s(X_1)} < \mathrm{s(X_2)} < \mathrm{s(X_3)} | \mathrm{Y_1} = 1, \mathrm{Y_2} = 2, \mathrm{Y_3} = 3\right\} \\ + \frac{1}{2}\mathbb{P}\left\{s(X_1) = s(X_2) < s(X_3) | Y_1 = 1, Y_2 = 2, Y_3 = 3\right\} \\ + \frac{1}{2}\mathbb{P}\left\{s(X_1) < s(X_2) = s(X_3) | Y_1 = 1, Y_2 = 2, Y_3 = 3\right\} \\ + \frac{1}{6}\mathbb{P}\left\{s(X_1) = s(X_2) = s(X_3) | Y_1 = 1, Y_2 = 2, Y_3 = 3\right\},\end{aligned}$$

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 $s(X)$'s distribution is continuous, the last three terms in the term on the right hand side vanishes 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 function $s(x)$ ranks them in the right order.

Finally, observe that, when Assumption 1 is not fulfilled, the (scalar) VUS criterion can still be used to set the goal of ranking. However, the interpretation of optimal pre-orders (those with maximum VUS) in such a case becomes highly questionable, insofar as the latter can be very different. For instance, in the situation described in Example 21, 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 attained by the scoring functions 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.

4 The multi-class ranking problem

The goal is to learn from a sample $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ of i.i.d. copies of the random pair (X, Y) how to build a scoring function $\widehat{s}_n(x)$ of which ROC surface is as close as possible to ROC^*. Distance between ROC surfaces can be naturally

considered in many ways. Focus is here on an important example, the distance related to the L_1-norm:

$$d_1(s,s') = \int\int_{(\alpha,\gamma)\in[0,1]^2} |\mathrm{ROC}(s,\alpha,\gamma) - \mathrm{ROC}(s',\alpha,\gamma)| d\alpha d\gamma.$$

Attention of the reader should be drawn to the fact that the notation $d_1(s,s')$ used above for convenience represents by no means a distance between the scoring functions s and s', but between their ROC surfaces. Having equipped the ROC space $[0,1]^3$ with this metric topology, we introduce the corresponding notion of *ranking consistency*.

Definition 41. (Ranking consistency) *Suppose that Assumption 1 is fulfilled. Let* $\{s_n\}$ *be a sequence of scoring functions on* $\mathcal{X}$*. It is said* VUS*-consistent (respectively, strongly* VUS*-consistent), when* $d_1(s_n, s^*) \to 0$ *in probability (respectively, with probability one) as* n *tends to infinity.*

By virtue of Theorem 32, under Assumption 1, $d_1(s,s^*)$ can be expressed as the *deficit of volume under the* ROC *surface*, $\mathrm{VUS}^* - \mathrm{VUS}(s)$ namely, for any $(s,s^*) \in \mathcal{S}\times\mathcal{S}^*$, so that minimizing $d_1(s_n,s^*)$ obviously boils down to maximizing VUS(s). Of course, VUS maximization could be considered in absence of Assumption 1. However, one can easily see that VUS maximizers may then correspond to quite different rankings in some cases and their interpretation as optimal ranking rules becomes arguable.

Other metrics on the ROC space could be considered, such as that related to the L_∞-norm,

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

leading to a stronger notion of ranking consistency. The study of accuracy of ranking methods in this sense is beyond the scope of the present paper (in contrast to the L_1 situation, 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 [CV09e] and [CV09d]), for which consistency in sup-norm is guaranteed under some specific assumptions, will be considered in a forthcoming article.

The next subsections are devoted to the description of possible strategies for building consistent ranking rules.

4.1 K-partite ranking as a superposition of $K-1$ bipartite ranking tasks

In continuity with the way we have defined optimal scoring rules in the 3-partite setting in section 2, the three-class ranking problem can be viewed as a double bipartite ranking problem (and, more generally, K-partite ranking can be reduced to the simultaneous solving of $(K-1)$ bipartite ranking problems). The straightforward analysis below provides a simple theoretical formulation of this fact.

Theorem 41. (Deficit of VUS) *Suppose that Assumption 1 is fulfilled. Then, for any function* $s \in \mathcal{S}$*, we have*

$$\mathrm{VUS}^* - \mathrm{VUS}(s) \leq \left(\mathrm{AUC}^*_{F_1,F_2} - \mathrm{AUC}_{F_1,F_2}(s)\right) + \left(\mathrm{AUC}^*_{F_2,F_3} - \mathrm{AUC}_{F_2,F_3}(s)\right).$$

By a straightforward symmetry argument, it results from the theorem above that the deficit of VUS of any scoring function s can also be bounded by the quantity $(2/3)\sum_{1\leq k<l\leq 3}\{\mathrm{AUC}^*_{\mathrm{F_k,F_l}} - \mathrm{AUC}_{\mathrm{F_k,F_l}}(\mathrm{s})\}$. More importantly, this shows that a sequence of scoring functions 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 42. (Superposition of bipartite ranking problems) *Suppose that Assumption 1 is fulfilled. Let $\{s_n\}_{n\in\mathbb{N}}$ be a sequence of scoring functions. The following assertions are equivalent.*

(i) The sequence $\{s_n\}$ of scoring functions is asymptotically optimal (respectively, strongly asymptotically optimal) with respect to the VUS *criterion,* $\mathrm{VUS}(\mathrm{s_n}) \to \mathrm{VUS}^*$ *as* $\mathrm{n}\to\infty$.

(ii) As n goes to infinity, we simultaneously have $\mathrm{AUC}_{\mathrm{F_1,F_2}}(\mathrm{s_n}) \to \mathrm{AUC}^*_{\mathrm{F_1,F_2}}$ *and* $\mathrm{AUC}_{\mathrm{F_2,F_3}}(\mathrm{s_n}) \to \mathrm{AUC}^*_{\mathrm{F_2,F_3}}$.

The goal is thus to build a scoring function $\widehat{s}_n(x)$ based on training data with nearly optimal ROC curves $\mathrm{ROC}_{\mathrm{F_1,F_2}}(\widehat{\mathrm{s}}_\mathrm{n},.)$ and $\mathrm{ROC}_{\mathrm{F_2,F_3}}(\widehat{\mathrm{s}}_\mathrm{n},.)$ both at the same time. Hence, 3-class ranking can be cast in terms of a double-criteria optimization task, consisting in finding a scoring function s that simultaneously maximizes $\mathrm{AUC}_{\mathrm{F_1,F_2}}(\mathrm{s})$ and $\mathrm{AUC}_{\mathrm{F_2,F_3}}(\mathrm{s})$. Based on this observation, there are two successive stages in the approach we develop subsequently. The first step lies in solving each bipartite ranking separately, producing two scoring functions. Based on the latter, the second one then consists in computing a *median scoring rule*, the median being taken in a certain sense, which permits to preserve the asymptotic consistency properties of the original scoring functions. An algorithmic description of this procedure together with theoretical grounds for its validity are given in the next section. We point out that the idea of decomposing the K-partite ranking into several bipartite ranking subproblems has also been considered in [FHV09] (with a quite different way of performing the aggregation stage however, see section 5 below).

In the remainder of the present section, we briefly review two alternative methods for building consistent scoring functions in the general K-partite framework. Again, for simplicity's sake, they are described in the case $K = 3$ solely.

4.2 Plug-in scoring rule

As shown by Theorem 23, when Assumption 1 is fulfilled, the regression function $\eta(x)$ is an optimal scoring function. The so-termed *plug-in* approach consists of estimating the latter and use the resulting estimate as a scoring function. For instance, one may estimate the distribution *a posteriori* $(\eta_1(x),\ \ldots,\ \eta_K(x))$ by an empirical counterpart $(\widehat{\eta}_1(x),\ \ldots,\ \widehat{\eta}_K(x))$ based on the training data and consider the preorder on $\mathcal{X}$ induced by the estimator $\widehat{\eta}(x) = \sum_{k=1}^K k\cdot\widehat{\eta}_k(x)$, see [CV09c] and [CR11] for preliminary theoretical results based on this strategy in the bipartite context and [AT07] for an account of the plug-in approach in binary classification. It is hence 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 $\widehat{\eta}(x)$ to its $L_1(\mu)$-error as an estimate of $\eta(x)$, in the case where all class distributions have the same support for simplicity's sake (extension to the general situation is left to the reader).

Proposition 43. (Deficit of VUS (bis)) *Suppose that Assumption 1 is fulfilled. Let* $\widehat{\eta}$ *be an approximant of* $\eta(x)$*. Assume that both the random variables* $\eta(X)$ *and* $\widehat{\eta}(X)$ *are continuous. We have:* $\forall s \in \mathcal{S}$*,*

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

This result reveals that a $L_1(\mu)$-consistent estimator, *i.e.* an estimator $\widehat{\eta}_n$ such that $\mathbb{E}[|\eta(X) - \widehat{\eta}_n(X)|] = \int |\eta(x) - \widehat{\eta}(x)|\mu(dx)$ 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 with difficulties related to the curse of dimensionality.

4.3 Empirical VUS maximization

Exploiting the fact that the target $\mathcal{S}^*$ coincides with the set of scoring functions with maximum VUS,

$$\mathcal{S}^* = \arg\max_{s \in \mathcal{S}} \mathrm{VUS(s)},$$

a standard approach, based on the *empirical risk minimization* paradigm, see [Vap99], lies in optimizing a statistical counterpart of the unknown functional VUS(.) over a set $\mathcal{S}_1 \subset \mathcal{S}$ of scoring function candidates of quantifiable complexity (a *VC* major class of functions with finite *VC* dimension for instance, see [Dud99]). Based on the training dataset $\mathcal{D}_n$, a natural empirical counterpart of VUS(s), $s \in \mathcal{S}$, is the three-sample U-statistic

$$\widehat{\mathrm{VUS}}_n(s) = \frac{1}{n_1 n_2 n_3} \sum_{1 \leq i,\, j,\, k \leq n} h_s(X_i, X_j, X_k) \cdot \mathbb{I}\{Y_i = 1,\ Y_j = 2,\ Y_k = 3\}, \qquad (4)$$

with kernel given by

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

for any $(x_1, x_2, x_3) \in \mathcal{X}^3$. Computational complexity of empirical VUS calculation is investigated in [WBB08a].

A scoring function s maximizing $\widehat{\mathrm{VUS}}_n(.)$ over $\mathcal{S}_1$ is expected to approximately maximize VUS(.), when $\mathcal{S}_1$ is "rich enough" (when it contains an approximate maximizer of VUS(.), namely). Properties of the empirical VUS maximizer (bounds on its deficit of VUS, in particular) can be then easily investigated using 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 $\mathcal{S}_1$, following in the footsteps of [CLV08] in the bipartite 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 [CV09e] or [CV09d] requires, for instance, significant methodological progress.
Finally, we point out that a theoritical study of the empirical risk minimization strategy (ERM) in the K-partite ranking context has been carried out in [RA05], where a

different accuracy measure is used, 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 this case, ERM is shown to boil down to maximizing a weigthed sum of empirical AUC's and generalization bounds have been established for classes of scoring rules s with finite *k-Partite rank-shatter coefficient*.

5 Aggregating bipartite ranking rules: the Kendall consensus

In multiple-class classification, one may easily build predictive rules through solving a collection of binary classification rules, using either the "*one against one*" approach or the "*one versus all*" procedure (see [ASS01], [HT98], [VA99], [DTT04], [DB95], [BLZ05], [BDH$^+$05] and the references therein for instance) and proceeding then to a majority vote. Whereas it straightforward to give sense to the notion of majority prediction in the classification context, it is not that easy when the issue is to predict how to sort all $\mathcal{X}$'s elements. Aggregating binary relationships such as total preorders is an old issue, sending us back to the pioneer work of Condorcet in Social Choice theory, see [BB81]. Here we revisit this important problem from the multi-class ranking perspective. Our goal is to define a notion of "ranking consensus" that permits to build a VUS-consistent scoring rule based on $K-1$ scoring functions $s_n^{(1)}, \ \ldots, \ s_n^{(K-1)}$, when $s_n^{(k)}(x)$ is AUC-consistent for the bipartite ranking problem related to the pair $(F_k(dx), \ F_{k+1}(dx))$, for $1 \leq k < K$. Throughout this section, we assume that the F_k's are all absolutely continuous with respect to each other (in particular, $\mathcal{X} = \mathcal{X}_1 = \cdots = \mathcal{X}_K$), in order to exclude situations where some populations are easily separable.

5.1 On Kendall measure of agreement between scoring functions

Here, we shall say that two scoring functions $s_1(x)$ and $s_2(x)$ on $\mathcal{X}$ "agree" when the random variables $s_1(X)$ and $s_2(X)$ tend to increase or decrease together. A natural way of quantifying *agreement* is thus to consider the (theoretical) Kendall τ related to the pair $(s_1(X), s_2(X))$.
For clarity, we recall that the Kendall τ related to a pair (V, W) of real-valued random variables defined on the same probability space is given by:

$$\tau(V,W) = \mathbb{P}\left\{\left(V - V'\right) \cdot \left(W - W'\right) > 0\right\} + \frac{1}{2}\mathbb{P}\left\{V \neq V', \ W = W'\right\} + \frac{1}{2}\mathbb{P}\left\{V = V', \ W \neq W'\right\}.$$

The quantity $\tau(V, W)$ ranges from 0 (full disagreement) to 1 (full agreement). The empirical version, based on a sample of $N \geq 2$ independent copies $(V_1, W_1), \ \ldots, \ (V_N, W_N)$ of the pair (V, W), known as the Kendall τ statistic, is the U-statistic of degree 2

$$\widehat{\tau}_N = \frac{2}{N(N-1)} \sum_{1 \leq n < m \leq N} U\left((V_n, W_n), (V_m, W_m)\right),$$

with kernel given by

$$U\left((v,w),(v',w')\right) = \mathbb{I}\left\{\left(v - v'\right) \cdot \left(w - w'\right) > 0\right\} + \frac{1}{2}\mathbb{I}\left\{v = v', \ w \neq w'\right\} + \frac{1}{2}\mathbb{I}\left\{v \neq v', \ w = w'\right\},$$

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

The quantity $\tau_\nu(s_1, s_2) \stackrel{def}{=} \tau(s_1(X), s_2(X))$ thus measures to which extent two real-valued scoring functions s_1 and s_2 on a space $\mathcal{X}$ rank pairs of independent copies of a r.v. X drawn from a distribution ν on $\mathcal{X}$. Notice that $d_{\tau_\nu}(.,.) = (1 - \tau_\nu(.,.))/2$ defines a pseudo-metric on the set of real-valued scoring functions on $\mathcal{X}$ (it is positive, symmetric and the triangular inequality holds true). In addition, when ν's support coincides with $\mathcal{X}$, it is a metric on the set of preorders $\preceq_s$ defined by the elements s of $\mathcal{S}_1$. The following proposition shows that the AUC deviation between two scoring functions is controlled by the related probabilistic Kendall τ in a very simple fashion. It is essentially for this reason that the Kendall τ criterion plays a crucial role in the aggregation procedure we shall subsequently describe and analyze. One may refer to [BFB09] for efficient computation of Kendall τ statistics.

Proposition 51. *(*AUC AND KENDALL τ*) Let p be a real number in $(0, 1)$. Consider two probability distributions $F_1(dx)$ and $F_2(dx)$ on the set $\mathcal{X}$. Set $\nu(dx) = (1 - p)F_1(dx) + pF_2(dx)$. For any real-valued scoring functions $s_1(x)$ and $s_2(x)$ on $\mathcal{X}$, we have:*

$$|\mathrm{AUC}_{\mathrm{F}_1,\mathrm{F}_2}(\mathrm{s}_1) - \mathrm{AUC}_{\mathrm{F}_1,\mathrm{F}_2}(\mathrm{s}_2)| \leq \frac{1 - \tau_\nu\left(\mathrm{s}_1, \mathrm{s}_2\right)}{4\mathrm{p}(1-\mathrm{p})} = \frac{\mathrm{d}_{\tau_\nu}(\mathrm{s}_1, \mathrm{s}_2)}{2\mathrm{p}(1-\mathrm{p})}.$$

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 [CLV08] is fulfilled.

Proposition 52. (AUC AND KENDALL τ (BIS)) *Consider two probability distributions $F_1(dx)$ and $F_2(dx)$ on the set $\mathcal{X}$, absolutely continuous with respect to each other. Let (X, Y) be a pair of random variables valued in $\mathcal{X} \times \{1, 2\}$ and such that $\mathbb{P}\{Y = 2\} = 1 - \mathbb{P}\{Y = 1\} = p \in (0, 1)$ and X's conditional distribution given $Y = k$ is $F_k(dx)$, $k = 1,\ 2$. Assume that the r.v. $\zeta(X) = \mathbb{P}\{Y = 2 \mid X\}$ is continuous and there exist $c < \infty$ and $a \in (0, 1)$ such that*

$$\forall x \in \mathcal{X},\ \ \mathbb{E}\left[|\zeta(X) - \zeta(x)|^{-a}\right] \leq c. \tag{5}$$

Then, we have for all pair of real valued scoring functions $(s, s^*) \in \mathcal{S} \times \mathcal{S}^*_{1,2}$,

$$d_{\tau_\nu}(s^*, s) \leq C \cdot \left(\mathrm{AUC}^*_{\mathrm{F}_1,\mathrm{F}_2} - \mathrm{AUC}_{\mathrm{F}_1,\mathrm{F}_2}(\mathrm{s})\right)^{a/(1+a)},$$

with $C = (3/2) \cdot c^{1/(1+a)} \cdot (2p(1-p))^{a/(1+a)}$.

Remark 51. (ON THE NOISE CONDITION.) *Recall that condition (5) is rather weak. Indeed, it is fulfilled for any $a \in (0, 1)$ as soon the distribution of the r.v. $\zeta(X)$ has a bounded density, see Corollary 8 in [CLV08].*

5.2 Aggregating solutions of bipartite ranking problems

We now introduce the following notion of *scoring function consensus*, based on the Kendall tau distance d_{τ_μ}.

Definition 51. (MEDIAN SCORING FUNCTION) *Let $\mathcal{S}_1 \subset \mathcal{S}_0$ be a set of real-valued scoring function candidates and $s^{(1)}, \ldots, s^{(K-1)}$ be $N \geq 1$ real-valued scoring functions. A median scoring function related to the collection $\{s^{(k)} : 1 \leq k < K\}$ and the set $\mathcal{S}_1$ is any scoring function $\overline{s} \in \mathcal{S}_0$ such that:*

$$\sum_{k=1}^{K-1} \tau_\mu \left(\overline{s}, s^{(k)} \right) = \sup_{s \in \mathcal{S}_1} \sum_{k=1}^{K-1} \tau_\mu \left(s, s^{(k)} \right). \tag{6}$$

It should be pointed up that, in general the supremum appearing on the right hand side of Eq. (6) 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 is not necessarily unique however). In particular, this is the case when considering *piecewise constant scoring functions* such as those produced by the bipartite ranking algorithms proposed in [CDV11], [CV09d], [CV09c] or [CN09], see section 6 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. Its proof is obvious and thus omitted.

Proposition 53. *Let $s^{(k)}$ be a real-valued scoring function in $\mathcal{S}^*_{k,k+1}$ for $1 \leq k < K$. Suppose that for all $(k,l) \in \{1, \ldots, K-1\}^2$, for all $(x, x') \in \mathcal{X}^2$, we have:*

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

Then,

$$\inf_{s \in \mathcal{S}_0} \sum_{k=1}^{K-1} d_{\tau_\mu} \left(s, s^{(k)} \right) = 0 \tag{7}$$

and

$$\mathcal{S}_0^* = \arg\inf_{s \in \mathcal{S}_0} \sum_{k=1}^{K-1} d_{\tau_\mu} \left(s, s^{(k)} \right),$$

where $\mathcal{S}_0^* = \mathcal{S}_0 \cap \mathcal{S}^*$ denotes the set of real-valued optimal scoring functions.

The proposition above reveals that "consensus scoring functions", in the sense of Definition 51, based on $K-1$ optimal scoring functions are still optimal solutions for the global multi-class ranking problem and that, conversely, elements of $\mathcal{S}_0^*$ necessarily achieve the supremum (7). This naturally suggests to implement the following two-stage procedure, that consists in 1) solving the bipartite ranking subproblem related to the pair (F_k, F_{k+1}) of consecutive class distributions, yielding a scoring function $s^{(k)}$, for $1 \leq k < K$ and 2) computing a median (6), 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 [CV09e] 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 in terms of VUS.

5.2.1 The aggregation procedure

Now the rationale behind the way the multi-class ranking problem can be divided into a series of bipartite subproblems, we recapitulate the successive steps of the "pairwise aggregation" approach. Based on two independent samples, a sample $\mathcal{D} = \{(X_i, Y_i) : 1 \leq i \leq n\}$ with i.i.d. labeled observations and $\mathcal{D}' = \{X_i', : \ 1 \leq i \leq n'\}$ a sample with unlabeled observations (the first one is used for training bipartite ranking rules, while one uses the other for consensus calculation), it is implemented in two steps, provided a bipartite ranking algorithm $\mathcal{A}$ is available, as follows.

The Kendall aggregation approach for multi-class ranking

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

1. **A series of bipartite ranking tasks.** For $k = 1, \ldots, K-1$, run algorithm $\mathcal{A}$ in order to train a scoring function $\widehat{s}^{(k)}(x)$ based on the truncated samples $\mathcal{D}_k$ and $\mathcal{D}_{k+1}$ of observations in $\mathcal{D}$ with labels k and $k+1$ respectively.
2. **Aggregating scoring rules.** Compute $\widehat{s}(x)$ in $\mathcal{S}_1$ such that:

$$\sum_{k=1}^{K-1} \tau_{\widehat{\mu}}\left(\widehat{s}, \widehat{s}^{(k)}\right) = \max_{s \in \mathcal{S}_1} \sum_{k=1}^{K-1} \tau_{\widehat{\mu}}\left(s, \widehat{s}^{(k)}\right),$$

where $\widehat{\mu}(dx)$ denotes the empirical distribution computed using the (unlabeled) sample $\mathcal{D}'$.

Before stating an asymptotic result providing theoretical grounds for this aggregation method, a few remarks are in order.

We first underline that the use of two independent samples, one for the bipartite ranking tasks and the other for the median computation, can be seen as an adaptation of the well-known "two-sample trick", widely used in semi-parametric statistics, to avoid possibly harmful dependencies, permitting thus to study the asymptotic behavior of the (functional) statistic $\widehat{s}(x)$ constructed in two stages. However, according to our experience, using a single (and thus larger) data set for both stages yields similar, or even better, numerical results in practice.

Rank prediction *vs.* scoring rule learning. When the goal is to rank accurately new unlabeled datasets, rather than to learn a nearly optimal scoring function explicitly, the following variant of the procedure described above can be considered, avoiding in particular the (difficult) optimization stage over a set of scoring functions. 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 $\mathcal{D}_X$'s elements, one may aggregate their restrictions to the finite set $\mathcal{D}_X \subset \mathcal{X}$, or the ranks of the unlabeled data as defined by the $s^{(k)}$'s, more simply.

Practical implementation. 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 [MPPB07], [FKM$^+$03] or [LL03] for instance. As shown in [Hud08] or [Wak98] 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 (6) 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 [BGH89],[CH98], [LMC99] or [MM09] 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.

5.2.2 Main result

The following theorem reveals that the notion of median introduced in Definition 51 preserves AUC consistency and thus yields a VUS consistent ranking when based on $K-1$ rankings, each being AUC-consistent for the bipartite subproblem related to a specific pair of class distributions (F_k, F_{k+1}), $1 \leq k < K$. For simplicity's sake again, it is stated in the 3-class setting.

Theorem 54. *Suppose that assumptions of Proposition 53 are satisfied. Let $\mathcal{S}_1 \subset \mathcal{S}_0$ be some set of real-valued scoring functions such that $\mathcal{S}^* \cap \mathcal{S}_1 \neq \emptyset$. Let $s_n(x)$ and $s'_n(x)$ be* AUC*-consistent sequences of scorings functions in $\mathcal{S}_1$ for the bipartite ranking problems related to the pairs of distributions (F_1, F_2) and (F_2, F_3) respectively. If there exists a median scoring rule $\overline{s}_n(x)$ in the sense of Definition 51, it is then* VUS*-consistent.*

The result stated above deserves some comments. It provides a theoretical basis for the ranking method proposed above that reduces the K-partite task to a series of bipartite tasks (in this respect, notice incidentally that even proving consistency of classification methods that reduce multi-class to binary is far from obvious in general, see [ASS01]) and its goal is to explain the main idea rather than to give the results in full generality. Of course, it is not realistic to assume that $\mathcal{S}_1$ contains some optimal scoring functions. However, through careful examination of its proof, one can see that this simplifying hypothesis can easily be replaced by the assumption that the scoring functions s_n, s'_n and $\overline{s}_n(x)$ 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 $\mathrm{VUS}(\mathrm{s}_\mathrm{n}^*) \to \mathrm{VUS}^*$ as $n \to \infty$, at the price of an additional bias term.

6 Illustrative numerical experiments

It is the purpose of this section to illustrate the approach described above by simulation 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 31). 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 [CDV11] for a detailed description of the algorithm (as well as [CV09e] for rigorous statistical foundations of this method). Precisely, we used a package for R statistical software (see http://www.r-project.com) implementing TreeRank (with the "default" parameters: minsplit = (size of training sample)/20, maxdepth = 10, mincrit = 0), available at http://treerank.sourceforge.net, see [BCDV09]. The scoring rules produced at stage 1 are thus (tree-structured and) piecewise constant, making the aggregating procedure described in sub-subsection 5.2.1 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 subsection 5.2, 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.

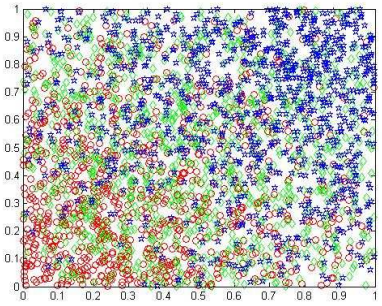
For comparison purpose, we have also implemented two ranking algorithms, RankBoost (when aggregating 30 stumps, see [RCMS05]) and SVMRank (with linear and Gaussian kernels with repective parameters $C = 20$ and $(C, \gamma) = (0.01)$, see [HGO00]), using the SVM-light implementation available at http://svmlight.joachims.org/. We have also used the RankRLS method (http://www.tucs.fi/RLScore, see [PTA$^+$07]) that implements a regularized least square algorithm with linear kernel ("$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:

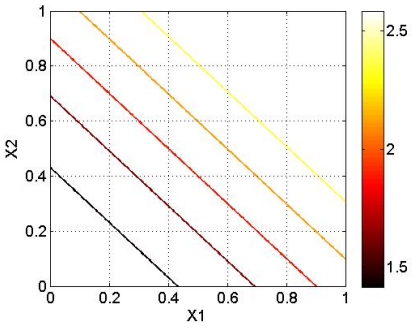
$$F_1(dx) = \mathcal{N}_{[0,1]^2}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1/4 & 0 \\ 0 & 1/4 \end{pmatrix}\right)$$

$$F_2(dx) = \mathcal{N}_{[0,1]^2}\left(\begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}, \begin{pmatrix} 1/4 & 0 \\ 0 & 1/4 \end{pmatrix}\right)$$

$$F_3(dx) = \mathcal{N}_{[0,1]^2}\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1/4 & 0 \\ 0 & 1/4 \end{pmatrix}\right)$$



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

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

Method	$\overline{\text{VUS}}\ (\widehat{\sigma})$
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)

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 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 $s^{(3)}(x)$ based on data with labels "1" and "3". Finally, $s^{(1)}$ and $s^{(2)}$ are aggregated through the procedure described in sub-subsection 5.2.1, 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 "RankBoost-VUS", 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

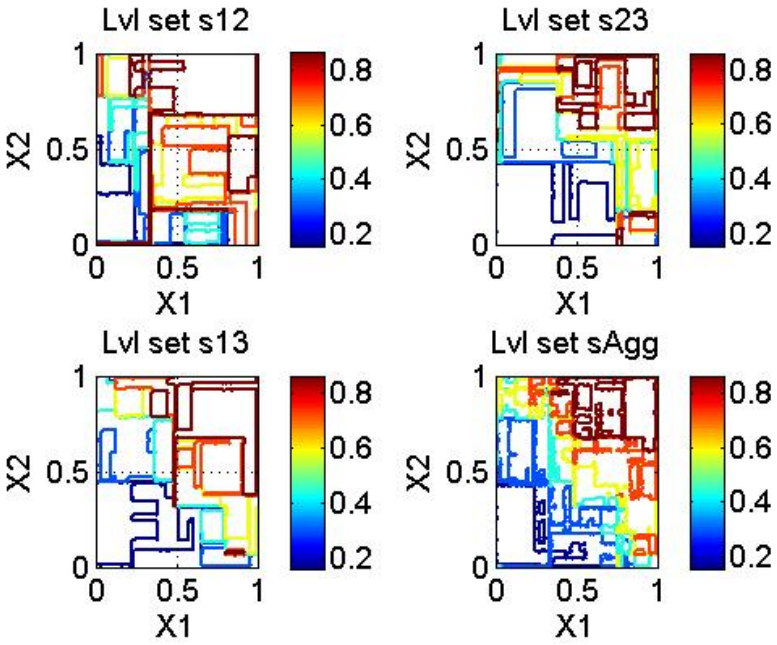


Fig. 4 Levels sets of the scoring functions "TreeRank 1v2", "TreeRank 2v3", "TreeRank 1v3" and "TreeRank Agg" in a top-down left-right manner

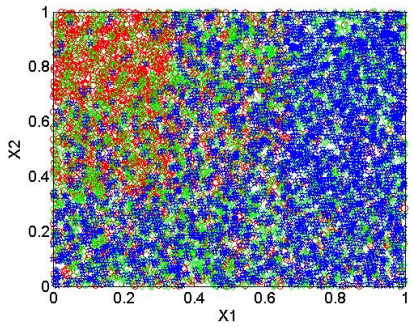
lin" and "RLScore gauss"). Averages ($\overline{\mathrm{VUS}}$) over the 50 training samples have been next computed, as well as standard deviations $\widehat{\sigma}$, they are given in Table 1 with the results of the earlier described algorithms. For comparison purpose, some level sets of the TreeRank 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 5.a 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^* as the function constant on each of these squares depicted by Fig. 5.b). 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 $\phi_{2,3} = 1.3 s^*_{2,3}(x)$. 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 Proposition 22, 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.

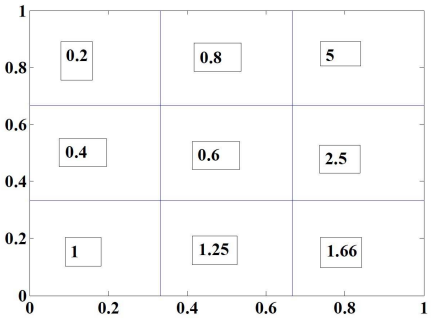
Here 50 training samples of size $n = 9000$ plus a test sample of size 9000 have been generated. The performance results are reported in Table 3. For comparison purpose, some level sets of the scoring function learnt on the first training sample for each method is represented in Fig. 6.

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

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

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

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

An example based on real data. We finally illustrate the methodology promoted in this paper by implementing it on a real data set, the *Cardiotocography Data Set* considered in [FA10] 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, depending on the degree of anomaly observed: 1 for "normal", 2 for "suspect" and 3 for "pathologic".

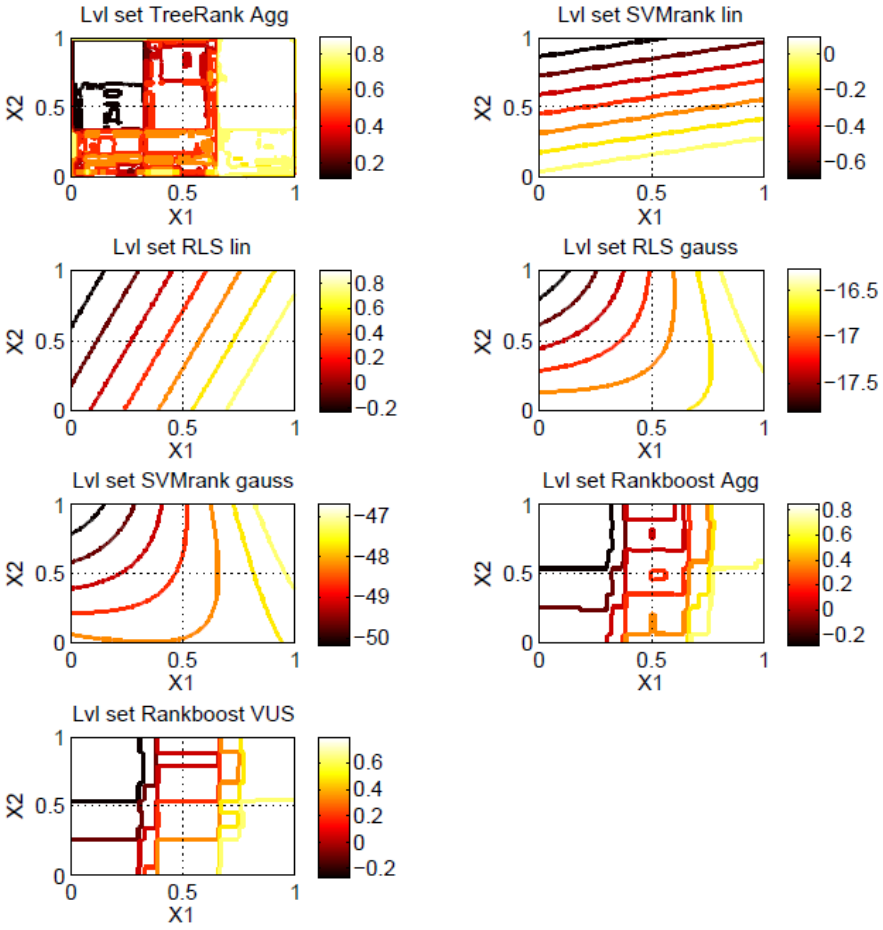


Fig. 6 Levels sets of the scoring functions "TreeRank Agg", "SVMrank lin", "RLScore lin", "RLScore gauss", "SVMrank gauss", "RankBoostVUS", "RankBoost Agg".

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

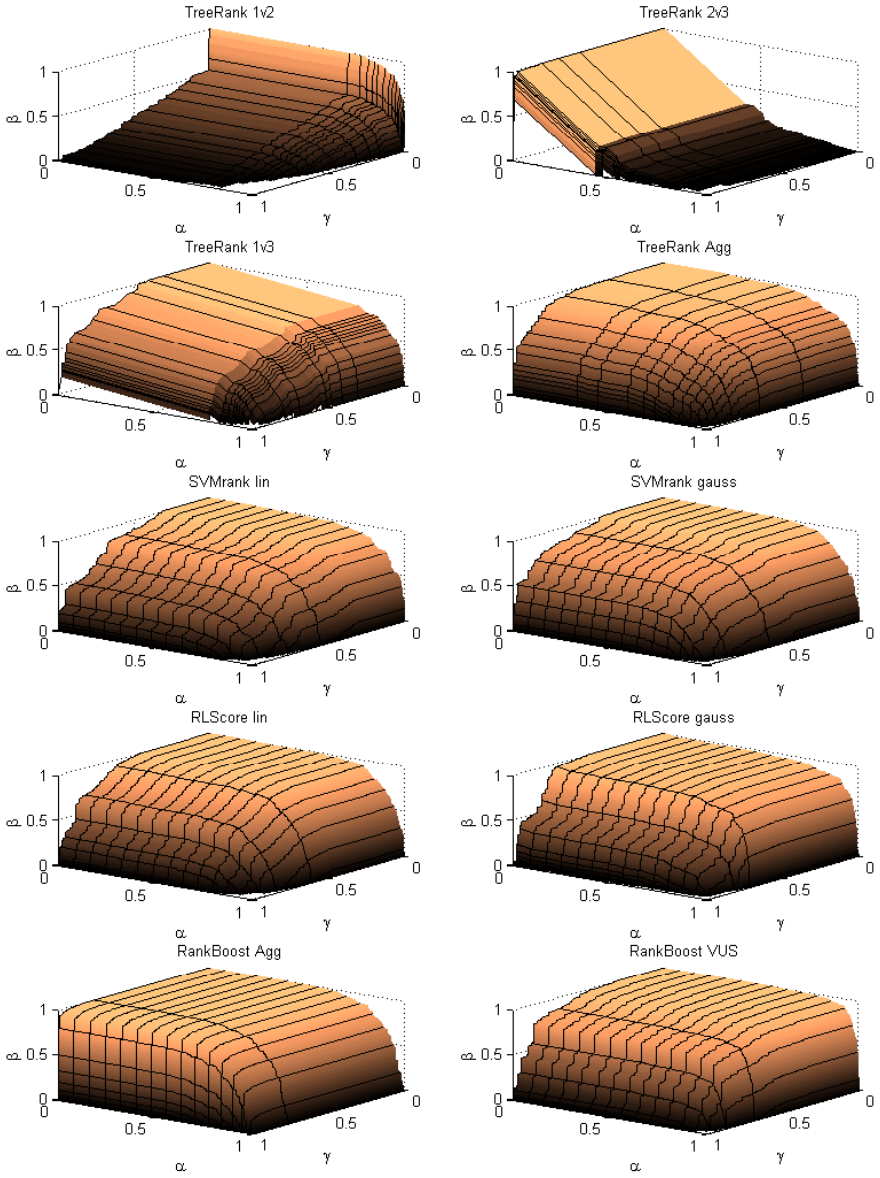


Fig. 7 ROC surfaces "test" of the scoring functions bipartite "TreeRank", "TreeRank Agg", "SVMrank lin", "RLScore lin", "RLScore gauss", "SVMrank gauss", "RankBoostVUS", "RankBoost Agg".

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

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.

These empirical results only aim at illustrating the Kendall aggregation approach for multi-class ranking, the limited goal pursued here being to show how aggregation helps to improve results. Beyond the theoretical validity framework sketched in section 5, 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 subsection 5.2. A complete and detailed empirical analysis of the merits and limitations of this procedure will be the subject of a forthcoming article, where comparisons with competitors will be carried out (based on real datasets in particular) and computational issues be discussed at length.

7 Conclusion

In this article, we have presented theoretical work on multi-class ranking. 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 regres-

sion function is proved to define an optimal ranking rule in this setting, highlighting the connection between K-partite ranking and ordinal regression. 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 multi-class 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. 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 [FHV09]. For this purpose, we have introduced a specific notion of *median scoring function* based on the (probabilistic) Kendall τ distance and shown that it 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 a dedicated article 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 [CV09e] and [CDV11], 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 - Technical Proofs

Proof of Theorem 23

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

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

The proof is based on the next lemma.

Lemma 1 *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* $j \in \{1, ..., K\}$, *we have*

$$\sum_{i=j}^{K} \eta_i(x) \leq \sum_{i=j}^{K} \eta_i(x'). \tag{8}$$

Additionally, a strict version of inequality (8) 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 $\Phi_{k,l}(x) = \dfrac{p_l \eta_k(x)}{p_k \eta_l(x)}$ and $\eta_l(x) = 1 - \sum_{i \neq l} \eta_i(x)$, we clearly have

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

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

$$\begin{aligned} \eta_m(x) &\leq \eta_m(x') + \sum_{i<j-1} \left\{ \eta_m(x)\eta_i(x') - \eta_m(x')\eta_i(x) \right\} + \sum_{i>j-1} \left\{ \eta_m(x)\eta_i(x') - \eta_m(x')\eta_i(x) \right\}, \\ &\leq \eta_m(x') + \sum_{i>j-1} \left\{ \eta_m(x)\eta_i(x') - \eta_m(x')\eta_i(x) \right\}. \end{aligned} \tag{9}$$

Summing up, term-by-term, inequalities (9) for $m = j, \ldots, K$, one gets that

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

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 (8) stated in Lemma 1 for $j = 1, \ldots, K$.

Proof of Theorem 32

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 $\mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_3}(\mathrm{s}^*, 1-\alpha) \geq \mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_3}(\mathrm{s}, 1-\alpha)$ for all $\alpha \in [0,1]$. Hence, as the desired bound obviously holds true on the set $\{(\alpha,\gamma) : \ \gamma > \mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_3}(\mathrm{s}^*, 1-\alpha)\} \subset \{(\alpha,\gamma) : \ \gamma > \mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_3}(\mathrm{s}, 1-\alpha)\}$, we place ourselves on the complementary set $\{(\alpha,\gamma) : \ \gamma \leq \mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_3}(\mathrm{s}^*, 1-\alpha)\}$, on which we have

$$\begin{aligned} \mathrm{ROC}(\mathrm{s}^*, \alpha, \gamma) - \mathrm{ROC}(\mathrm{s}, \alpha, \gamma) \leq &\left(\mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_2}(\mathrm{s}^*, 1-\alpha) - \mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_2}(\mathrm{s}, 1-\alpha)\right) + \\ &\left(\mathrm{ROC}_{\mathrm{F}_3,\mathrm{F}_2}(\mathrm{s}, \gamma) - \mathrm{ROC}_{\mathrm{F}_3,\mathrm{F}_2}(\mathrm{s}^*, \gamma)\right). \end{aligned}$$

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: $\mathrm{ROC}_{\mathrm{H},\mathrm{G}}(\mathrm{s}, \alpha) \leq \beta \Leftrightarrow \alpha \leq \mathrm{ROC}_{\mathrm{G},\mathrm{H}}(\mathrm{s}, \beta)$). The first part of the result is thus established.

Turning now to the second part, the bound stated obviously holds true on the set $\{(\alpha,\gamma) : \ \gamma > \mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_3}(\mathrm{s}^*, 1-\alpha)\}$.
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 $\{(\alpha,\gamma) : \ \gamma \leq \mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_3}(\mathrm{s}^*, 1-\alpha)\}$, we may then write:

$$\mathrm{ROC}(\mathrm{s}^*,\alpha,\gamma) - \mathrm{ROC}(\mathrm{s},\alpha,\gamma) \leq -\mathbb{E}[\mathrm{m}_1(\mathrm{X})|\mathrm{Y}=2] - \mathbb{E}[\mathrm{m}_3(\mathrm{X})|\mathrm{Y}=2].$$

Considering the first ROC curve deficit, we have:

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

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:

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

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

$$\frac{p_3}{p_2}\mathbb{E}[m_1(X)|Y=3] = \mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_3}(\mathrm{s},1-\alpha) - \mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_3}(\mathrm{s}^*,1-\alpha),$$

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:

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

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.

Proof of Theorem 33

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

$$\mathrm{ROC}(\mathrm{s}^*,\alpha,\gamma) \geq \mathrm{ROC}(\mathrm{s},\alpha,\gamma). \tag{10}$$

Observe that, if $\gamma > \mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_3}(\mathrm{s}^*,1-\alpha)$, this implies that $\gamma > \mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_3}(\mathrm{s},1-\alpha)$, 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. (10) with $\beta = 0$ (respectively, with $\alpha = 1$).

Proof of Theorem 41

Let $(s, s^*) \in \mathcal{S} \times \mathcal{S}^*$. Notice that, as $s^* \in \mathcal{S}^*_{1,3}$, we have $\{(\alpha,\gamma):\ \gamma \leq \mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_3}(\mathrm{s},1-\alpha)\} \subset \{(\alpha,\gamma):\ \gamma \leq \mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_3}(\mathrm{s}^*,1-\alpha)\}$, so that

$$\begin{aligned}
\mathrm{ROC}^*(\alpha,\gamma) - \mathrm{ROC}(\mathrm{s},\alpha,\gamma) &\leq \{\mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_2}(\mathrm{s}^*,1-\alpha) - \mathrm{ROC}_{\mathrm{F}_3,\mathrm{F}_2}(\mathrm{s}^*,\gamma) \\
&\quad - \big(\mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_2}(\mathrm{s},1-\alpha) - \mathrm{ROC}_{\mathrm{F}_3,\mathrm{F}_2}(\mathrm{s},\gamma)\big)_+\} \\
&\quad \times \mathbb{I}\{\gamma \leq \mathrm{ROC}^*_{\mathrm{F}_1,\mathrm{F}_3}(1-\alpha)\} \\
&\leq \{\mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_2}(\mathrm{s}^*,1-\alpha) - \mathrm{ROC}_{\mathrm{F}_3,\mathrm{F}_2}(\mathrm{s}^*,\gamma) \\
&\quad - \mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_2}(\mathrm{s},1-\alpha) - \mathrm{ROC}_{\mathrm{F}_3,\mathrm{F}_2}(\mathrm{s},\gamma)\} \\
&\quad \times \mathbb{I}\{\gamma \leq \mathrm{ROC}^*_{\mathrm{F}_1,\mathrm{F}_3}(1-\alpha)\} \\
&\leq \big(\mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_2}(\mathrm{s}^*,1-\alpha) - \mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_2}(\mathrm{s},1-\alpha)\big) \\
&\quad - \big(\mathrm{ROC}_{\mathrm{F}_3,\mathrm{F}_2}(\mathrm{s}^*,\gamma) - \mathrm{ROC}_{\mathrm{F}_3,\mathrm{F}_2}(\mathrm{s},\gamma)\big)\,.
\end{aligned}$$

Integrating over $(\alpha,\gamma) \in [0,1]^2$ then yields the desired bound, using the fact that, for any $s \in \mathcal{S}_0$, $\int_{\gamma=0}^1 \mathrm{ROC}_{\mathrm{F}_3,\mathrm{F}_2}(\mathrm{s},\gamma)\mathrm{d}\gamma = 1 - \mathrm{AUC}_{\mathrm{F}_2,\mathrm{F}_3}(\mathrm{s})$.

Proof of Proposition 43

By virtue of proposition 41, we have:

$$\mathrm{VUS}^* - \mathrm{VUS}(\widehat{\eta}) \leq \big(\mathrm{AUC}^*_{\mathrm{F}_1,\mathrm{F}_2} - \mathrm{AUC}_{\mathrm{F}_1,\mathrm{F}_2}(\widehat{\eta})\big) + \big(\mathrm{AUC}^*_{\mathrm{F}_2,\mathrm{F}_3} - \mathrm{AUC}_{\mathrm{F}_2,\mathrm{F}_3}(\widehat{\eta})\big)\,.$$

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

$$\mathrm{AUC}^*_{\mathrm{F}_1,\mathrm{F}_2} - \mathrm{AUC}_{\mathrm{F}_1,\mathrm{F}_2}(\widehat{\eta}) = \frac{1}{2\mathrm{p}_1\mathrm{p}_2}\mathbb{E}[|\eta_1(\mathrm{X})\eta_2(\mathrm{X}') - \eta_1(\mathrm{X}')\eta_2(\mathrm{X})| \cdot \mathbb{I}\{(\mathrm{X},\mathrm{X}') \in \Gamma\}],$$

where

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

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

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

Then, we get:

$$\mathrm{AUC}^*_{\mathrm{F}_1,\mathrm{F}_2} - \mathrm{AUC}_{\mathrm{F}_1,\mathrm{F}_2}(\widehat{\eta}) \leq \frac{1}{2\mathrm{p}_1\mathrm{p}_2}\mathbb{E}[|\eta(\mathrm{X}) - \eta(\mathrm{X}')|\mathbb{I}\{(\mathrm{X},\mathrm{X}') \in \Gamma\}].$$

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

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

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

Proof of Proposition 51

Recall that $\tau_\nu(s_1,s_2) = 1 - 2d_{\tau_\nu}(s_1,s_2)$, where $d_{\tau_\nu}(s_1,s_2)$ is given by:

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

Observe first that, for all $s \in \mathcal{S}_0$, $\mathrm{AUC}_{\mathrm{F}_1,\mathrm{F}_2}(\mathrm{s})$ may be written as:

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

Notice also that, using Jensen's inequality, one easily obtain that the quantity $2p(1-p)|\mathrm{AUC}_{\mathrm{F}_1,\mathrm{F}_2}(\mathrm{s}_1) - \mathrm{AUC}_{\mathrm{F}_1,\mathrm{F}_2}(\mathrm{s}_2)|$ is bounded by the expectation of the random variable

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

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

Proof of Proposition 52

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

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

Recall also that

$$2p(1-p)\left(\mathrm{AUC}^*_{\mathrm{F}_1,\mathrm{F}_2} - \mathrm{AUC}_{\mathrm{F}_1,\mathrm{F}_2}(\mathrm{s})\right) = \mathbb{E}\left[|\zeta(X)-\zeta(X')|\mathbb{I}\{(X,X') \in \Gamma_s\}\right] + \mathbb{P}\{s(X) = s(X'),\ (Y,Y') = (-1,+1)\},$$

see Example 1 in [CLV08] for instance.
Observe that Hölder inequality combined with the noise condition shows that the quantity $\mathbb{E}\left[\mathbb{I}\{(X,X') \in \Gamma_s\}\right]$ is bounded by

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

In addition, we have

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

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

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

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

Proof of Theorem 54

Let $(s_n^{(1)}, s_n^{(2)})$ be a sequence of real-valued scoring functions in $\mathcal{S}_1$ such that, as $n \to \infty$, $\mathrm{AUC}_{\mathrm{F_1,F_2}}(\mathrm{s_n^{(1)}}) \to \mathrm{AUC}^*_{\mathrm{F_1,F_2}}$ and $\mathrm{AUC}_{\mathrm{F_2,F_3}}(\mathrm{s_n^{(2)}}) \to \mathrm{AUC}^*_{\mathrm{F_2,F_3}}$. Here we consider the following consensus measure: $\forall s \in \mathcal{S}_1$,

$$\Delta_n(s) = d_{\tau_\mu}\left(s, s_n^{(1)}\right) + d_{\tau_\mu}\left(s, s_n^{(2)}\right).$$

Let $s^* \in \mathcal{S}_1 \cap \mathcal{S}^*$. With $\mu_{1,2} = (p_1/(1-p_3))F_1 + (p_2/(1-p_3))F_2$, Proposition 51, combined with the triangular inequality applied to the pseudo-distance $d_{\tau_{\mu_{1,2}}}$, implies that

$$\begin{aligned}
\mathrm{AUC}^*_{\mathrm{F_1,F_2}} - \mathrm{AUC}_{\mathrm{F_1,F_2}}(\bar{\mathrm{s}}_\mathrm{n}) &\leq \frac{d_{\tau_{\mu_{1,2}}}(s^*, \bar{s}_n)}{p_1 p_2/(1-p_3)^2} \\
&\leq \frac{d_{\tau_{\mu_{1,2}}}(\bar{s}_n^{(1)}, \bar{s}_n) + d_{\tau_{\mu_{1,2}}}(s^*, \bar{s}_n^{(1)})}{p_1 p_2/(1-p_3)^2} \\
&\leq \frac{d_{\tau_{\mu_{1,2}}}(s^*, \bar{s}_n^{(1)})}{p_1 p_2/(1-p_3)^2} + \frac{d_{\tau_\mu}(\bar{s}_n^{(1)}, \bar{s}_n)}{p_1 p_2}.
\end{aligned}$$

The desired result finally follows from Proposition 52 combined with the AUC-consistency assumptions.

References

[Aga08] S. Agarwal. Generalization bounds for some ordinal regression algorithms. In *Proceedings of the 19th International Conference on Algorithmic Learning Theory*, 2008.

[AGH+05] S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth. Generalization bounds for the Area Under the ROC Curve. *JMLR*, 6:393–425, 2005.

[ASS01] E. Allwein, R. Schapire, and Y. Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research*, (1):113–141, 2001.

[AT07] J.Y. Audibert and A.B. Tsybakov. Fast learning rates for plug-in classifiers. *Ann. Statist.*, 32:608–633, 2007.

[BB81] J.P. Barthélémy and B.Montjardet. The median procedure in cluster analysis and social choice theory. *Mathematical Social Sciences*, 1:235–267, 1981.

[BCDV09] N. Baskiotis, S. Clémençon, M. Depecker, and N. Vayatis. R-implementation of the TreeRank algorithm. *Submitted for publication*, 2009.

[BDH+05] A. Beygelzimer, V. Dani, T. Hayes, J. Langford, and B. Zadrozny. Reductions between classification tasks. In *Proceedings of ICML'05*, 2005.

[BFB09] M. Bansal and D. Fernandez-Baca. Computing distances between partial rankings. *Information Processing Letters*, 109:238–241, 2009.

[BGH89] J.P. Barthélemy, A. Guénoche, and O. Hudry. Median linear orders: heuristics and a branch and bound algorithm. *European Journal of Operational Research*, 42(3):313–325, 1989.

[BLZ05] A. Beygelzimer, J. Langford, and B. Zadrozny. Weighted one against all. In *Proceedings of AAAI'05*, 2005.

[CDV11] S. Clémençon, M. Depecker, and N. Vayatis. Adaptive partitioning schemes for bipartite ranking. *Machine Learning*, 43(1):31–69, 2011.

[CH98] I. Charon and O. Hudry. Lamarckian genetic algorithms applied to the aggregation of preferences. *Annals of Operations Research*, 80:281–297, 1998.

[CLV08] S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and empirical risk minimization of U-statistics. *Ann. Statist.*, 36(2):844–874, 2008.

[CN09] S. Clémençon and N.Vayatis. Adaptive estimation of the optimal roc curve and a bipartite ranking algorithm. In *Proceedings of ALT'09*, 2009.

[CR11] S. Clémençon and S. Robbiano. Minimax learning rates for bipartite ranking and plug-in rules. In *Proceedings of ICML 11*, 2011.

[CV07] S. Clémençon and N. Vayatis. Ranking the best instances. *Journal of Machine Learning Research*, 8:2671–2699, 2007.

[CV09a] S. Clémençon and N. Vayatis. Empirical performance maximization based on linear rank statistics. In *NIPS*, volume 3559 of *Lecture Notes in Computer Science*, pages 1–15. Springer, 2009.

[CV09b] S. Clémençon and N. Vayatis. Nonparametric estimation of the precision-recall curve. In *26-th International Conference in Machine Learning*, 2009.

[CV09c] S. Clémençon and N. Vayatis. On partitioning rules for bipartite ranking. In *Proceedings of AISTATS*, number 5, pages 97–104. JMLR: W&CP, 2009.

[CV09d] S. Clémençon and N. Vayatis. Overlaying classifiers: a practical approach to optimal scoring. *To appear in Constructive Approximation*, 2009.

[CV09e] S. Clémençon and N. Vayatis. Tree-based ranking methods. *IEEE Transactions on Information Theory*, 55(9):4316–4336, 2009.

[DB95] T. G. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Articial Intelligence Research*, (2):263–286, 1995.

[DOMB00] S. Dreiseitl, L. Ohno-Machado, and M. Binder. Comparing three-class diagnostic tests by three-way ROC analysis. *Medical Decision Making*, 20:323–331, 2000.

[DS66] D.M.Green and J.A. Swets. *Signal detection theory and psychophysics*. Wiley, 1966.

[DTT04] R. Debnath, N. Takahide, and H. Takahashi. A decision based one-against-one method for multi-class support vector machine. *Pattern Analysis and its Applications*, (7):164–175, 2004.

[Dud99] R.M. Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, 1999.

[EMK05] D.C. Edwards, C.E. Metz, and M.A. Kupinski. The hypervolume under the roc hypersurface of 'near-guessing' and 'near-perfect' observers in n-class classification tasks. *IEEE Trans. Med. Imag.*, 24:293–299., 2005.

[FÖ2] J. Fürnkranz. Round robin classification. *Journal of Machine Learning Research*, 2:721–747, 2002.

[FA10] A. Frank and A. Asuncion. UCI machine learning repository, 2010.

[Faw06] T. Fawcett. An Introduction to ROC Analysis. *Letters in Pattern Recognition*, 27(8):861–874, 2006.

[FE05] J.E. Fieldsend and R.M. Everson. Formulation and comparison of multi-class roc surfaces. In *Proceedings of ROCML 2005*, 2005.

[FE06] J.E. Fieldsend and R.M. Everson. Multi-class ROC analysis from a multi-objective optimisation perspective. *Pattern Recognition Letters*, 27:918–927, 2006.

[FHOS03] C. Ferri, J. Hernández-Orallo, and M.A. Salido. Volume under the roc surface for multi-class problems. In *Proc. of 14th European Conference on Machine Learning*, 2003.

[FHV09] J. Fürnkranz, E. Hüllermeier, and S. Vanderlooy. Binary decomposition methods for multipartite ranking. In *ECML PKDD '09*, 2009.

[FISS03] Y. Freund, R. D. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *JMLR*, 4:933–969, 2003.

[FKM+03] R. Fagin, R. Kumar, M. Mahdian, D. Sivakumar, and E. Vee. Comparing and aggregating rankings with ties. In *Proceedings of the 12-th WWW conference*, pages 366–375, 2003.

[Fla04] P. Flach. Tutorial: "the many faces of roc analysis in machine learning". part iii. Technical report, ICML'04, Alberta, Canada, 2004.

[HGO00] R. Herbrich, T. Graepel, and K. Obermayer. *Advances in Large Margin Classifiers*, chapter Large margin rank boundaries for ordinal regression, pages 115–132. MIT Press, 2000.

[HH09] J.C. Huhn and E. Hüllermeier. Is an ordinal class structure useful in classifier learning? *International Journal of Data Mining, Modelling and Management*, 1:45–67(23, 2009.

[HM82] J.A. Hanley and J. McNeil. The meaning and use of the area under a ROC curve. *Radiology*, (143):29–36, 1982.

[HT98] T. Hastie and R. Tibshirani. Classification by pairwise coupling. *The Annals of Statistics*, 26(2):451–471, 1998.

[HT01] D.J. Hand and R.J. Till. A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine Learning*, 45(2):171–186, 2001.

[Hud08] O. Hudry. NP-hardness results for the aggregation of linear orders into median orders. *Ann. Oper. Res.*, 163:63–88, 2008.

[KPWG01] S. Kramer, B. Pfahringer, G. Widmer, and M. De Groeve. Prediction of ordinal regression trees. *Fundamenta Informaticae*, 47:1001–1013, 2001.

[LD06] T. Landgrebe and R. Duin. 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*, November 2006.

[LF03] N. Lachiche and P.A. Flach. Improving accuracy and cost of two-class and multi-class probabilistic classifiers using roc curves. In *Proc. 20th International Conference on Machine Learning*, 2003.

[LL03] G. Lebanon and J. Lafferty. Conditional models on the ranking poset. In *Proceedings of NIPS'03*, 2003.

[LMC99] M. Laguna, R. Marti, and V. Campos. Intensification and diversification with elite tabu search solutions for the linear ordering problem. *Computers and Operations Research*, 26(12):1217–1230, 1999.

[LR05] E.L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer, 2005.

[LZ09] J. Li and X.H. Zhou. Nonparametric and semiparametric estimation of the three way receiver operating characteristic surface. *Journal of Statistical Planning and Inference*, 139:4133–4142, 2009.

[MM09] B. Mandhani and M. Meila. Tractable search for learning exponential models of rankings. In *Proceedings of AISTATS, Vol. 5 of JMLR:W&CP 5*, 2009.

[Mos99] D. Mossman. Three-way rocs. *Medical Decision Making*, 78:78–89, 1999.

[MPPB07] M. Meila, K. Phadnis, A. Patterson, and J. Bilmes. Consensus ranking under the exponential model. In *Conference on Artificial Intelligence (UAI)*, pages 729–734, 2007.

[NY04] C.T. Nakas and C.T. Yiannoutsos. Ordered multiple-class roc analysis with continuous measurements. *Statistics in Medicine*, Volume 23, Issue 22:3437–3449, 2004.

[Pep03] M. Pepe. *Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, 2003.

[PTA+07] T. Pahikkala, E. Tsivtsivadze, A. Airola, J. Boberg, and T. Salakoski. Learning to rank with pairwise regularized least-squares. In *Proceedings of SIGIR 2007 Workshop on Learning to Rank for Information Retrieval*, pages 27–33, 2007.

[QnL+10] J.R. Quevedo, E. Monta nés, O. Luaces, , and J.J. Del Coz. Adapting decision dags for multipartite ranking. In *ECML PKDD'10*, 2010.

[RA05] S. Rajaram and S. Agarwal. Generalization bounds for k-partite ranking. In *NIPS Workshop on Learning to Rank*, 2005.

[RCMS05] C. Rudin, C. Cortes, M. Mohri, and R. E. Schapire. Margin-based ranking and boosting meet in the middle. In P. Auer and R. Meir, editors, *Proceedings of COLT*, volume 3559 of *Lecture Notes in Computer Science*, pages 63–78. Springer, 2005.

[Rob10] S. Robbiano. Note on confidence regions for the ROC surface. Technical report, Telecom ParisTech, 2010.

[Rud06] C. Rudin. Ranking with a P-Norm Push. In *Proceedings of COLT*, 2006.

[Scu96] B.K. Scurfield. Multiple-event forced-choice tasks in the theory of signal detectability. *Journal of Mathematical Psychology*, 40:253–269, 1996.

[TDFMSN06] V. Torra, J. Domingo-Ferrer, J.M. Mateo-Sanz, and M. Ng. Regression for ordinal variables without underlying continuous variables. *Information Sciences*, 176:465–474, 2006.

[VA99] G. Venkatesan and S. Amit. Multiclass learning, boosting, and error-correcting codes. In *COLT '99: Proceedings of the twelfth annual conference on Computational learning theory*, pages 145–155, New York, NY, USA, 1999. ACM.

[Vap99] V.N. Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999, 1999.

[Wak98] Y. Wakabayashi. The complexity of computing medians of relations. *Resenhas*, 3(3):323–349, 1998.

[WB11] W. Waegeman and B. De Baets. On the era ranking representability of pairwise bipartite ranking functions. *Artif. Intell.*, 175:1223–1250, 2011.

[WBB08a] W. Waegeman, B. De Baets, and L. Boullart. On the scalability of ordered multi-class ROC analysis. *Computational Statistics and Data Analysis*, 52:3371–3388, 2008.

[WBB08b] W. Waegeman, B. De Baets, and L. Boullart. ROC analysis in ordinal regression learning. *Pattern Recognition Letters*, 29:1–9, 2008.