Mach Learn (2012) 88:433–444
DOI 10.1007/s10994-012-5290-y

TECHNICAL NOTE

ROC convex hull and nonparametric maximum likelihood estimation

Johan Lim · Joong-Ho Won

Received: 24 September 2011 / Accepted: 5 April 2012 / Published online: 4 May 2012

Abstract The ROC convex hull (ROCCH) is the least convex majorant of the empirical ROC curve, and represents the optimal ROC curve of a set of classifiers. This paper provides a probabilistic view to the ROCCH. We show that the ROCCH can be characterized as a nonparametric maximum likelihood estimator (NPMLE) of a convex ROC curve. We provide two NPMLE formulations, one unconditional and the other conditional, both of which yield the ROOCH as the solution. The solution technique relates the NPMLEs to convex optimization and classifier calibration. The connection between the NPMLEs and the ROCCH also suggests efficient algorithms to compute NPMLEs of a convex ROC curve, and a conditional bootstrap procedure for assessing uncertainties in the ROCCH.

Keywords ROC convex hull · ROC curve · Convexity · NPMLE · Geometric programming · Classifier calibration

1 Introduction

A receiver operating characteristic (ROC) curve is a graphical representation of two performance measures of binary classifiers, the false positive rate (FPR) and the true positive rate (TPR). The FPR is the probability of erroneously reporting negative instances as being positive, whereas the TPR is that of correctly reporting positive instances. The ROC space is a set of (FPR, TPR) pairs. Traditionally the ROC space is visualized by plotting the FPR on the x axis and the TPR on the y axis. A classifier that reports a class label corresponds to a point in the ROC space. To be specific, suppose a diagnostic test uses a continuous

Editor: Carla Brodley.

J. Lim
Department of Statistics, Seoul National University, Seoul, Korea
e-mail: johanlim@snu.ac.kr

J.-H. Won (✉)
VA Cooperative Studies Program, Mountain View, CA 94043, USA
e-mail: JoongHoJohann.Won@va.gov

variable X to diagnose a certain disease; if the value of X is larger than a critical value c, the subject of the diagnosis is classified into the disease (positive) class; otherwise into the non-disease (negative) class. Each critical value corresponds to a classifier. We use class conditional survival functions, defined as the complement of class conditional distribution functions, for notational convenience. Let S_{ND} and S_{D} be class conditional survival functions of X for the negative and the positive classes, respectively. That is, $S_i(c) = P_i(X > c)$ for $i = \mathrm{ND}, \mathrm{D}$, where P_i is a probability measure of X on class i. The FPR and the TPR of the given classifier are

$$\begin{aligned} \mathrm{FPR}(c) &= S_{\mathrm{ND}}(c), \\ \mathrm{TPR}(c) &= S_{\mathrm{D}}(c). \end{aligned} \tag{1}$$

The ROC curve plots $\mathrm{TPR}(c)$ against $\mathrm{FPR}(c)$ for all values of c. Explicitly, for $p \in [0, 1]$,

$$R(p) = S_{\mathrm{D}}\big(S_{\mathrm{ND}}^{-1}(p)\big) \tag{2}$$

by substituting $\mathrm{FPR}(c)$ with p and eliminating c.

A natural question that arises is how to estimate $R(p)$ from the observed data. If the TPR and the FPR are estimated by the empirical survival functions $\hat{S}_{\mathrm{D}}$ and $\hat{S}_{\mathrm{ND}}$, i.e., the proportions of the true positives and the false positives in the training data set, the estimated curve is a piecewise constant function called the empirical ROC curve. The empirical ROC curve is a nonparametric maximum likelihood estimator (NPMLE) of the $R(p)$. We may also impose geometrical constraints, such as convexity,[1] when estimating $R(p)$. Lloyd (2002) studies nonparametric and semiparametric maximum likelihood estimation of a convex ROC curve. Parametric methods enjoy the ability of producing a smooth, as well as convex, ROC curve. These methods assume that S_{ND} and S_{D} belong to a specific parametric family of distributions that guarantees (2) is convex. Pan and Metz (1997) and Metz and Pan (1999) use the normal error distribution, Dorfman et al. (1997) consider the gamma distribution, and Campbell and Ratnaparkhi (1993) introduce the Lomax family of curves.

The ROC curve of randomized diagnoses traces the least convex majorant (LCM) of the given ROC curve, well known as the ROC convex hull (ROCCH) to the machine learning community (Provost and Fawcett 2001). The properties and applications of the ROCCH have been extensively studied in the machine learning literature: Pareto optimality (Kim et al. 2006), repairing local non-convexity (Flach and Wu 2005), and cost-sensitive classification (Lim and Pyun 2009), to name a few. In particular, in the use of the ROCCH for classifier calibration, i.e., to transform classifier scores into posterior class probabilities, Fawcett and Niculescu-Mizil (2007) show that their ROCCH-based calibration method is equivalent to the pool-adjacent-violation (PAV) isotonic regression-based method by Zadrozny and Elkan (2002). However, its connection with maximum likelihood estimation has not been much explored, to the best of our knowledge.

In this paper, we show that the ROCCH is the nonparametric maximum likelihood estimator (NPMLE) of the true ROC curve when it is assumed convex. We formulate the NPMLE problem as a convex optimization problem, whose solution yields the ROCCH (Sect. 2). This convex programming formulation allows us to consider a conditional NPMLE, which also has the ROCCH as the optimal solution (Sect. 3). The benefit of this conditional NPMLE interpretation is that the uncertainty in the ROCCH can be systematically evaluated. To demonstrate this, we propose a conditional bootstrap procedure (Sect. 4).

[1] The use of the term ‘convex’ in the machine learning community in the context of ROC analysis is the opposite to its mathematical definition, as pointed out by Hand (2009). Since this article targets at the machine learning community, we adopt the machine learning convention.

2 Nonparametric maximum likelihood estimation of convex ROC curves

2.1 Likelihood

We can formulate the problem of NPMLE of ROC curves (under no constraint) using the class conditional survival functions (1) and the prior class probability that will be introduced shortly. Consider independent random samples from two classes, ND (negative) and D (positive). Let $x_{i1}, \dots, x_{in_i}$ be the observed diagnostic scores from class i whose survival function is S_i for $i = \mathrm{ND}, \mathrm{D}$, i.e., n_{ND} and n_{D} are the sizes of the negative and the positive classes, respectively. Complete observations occur on a subset of scores $x_1 < x_2 < \cdots < x_m$, where $\{x_1, \dots, x_m\}$ is the union of all observed scores $x_{i1}, \dots, x_{in_i}$ for $i = \mathrm{ND}, \mathrm{D}$. Note that this union partitions the axis of scores into $m + 1$ intervals. Assuming that the observations are mutually independent, the interval frequencies for each class follow a multinomial distribution (Metz et al. 1998). Then the likelihood of the observation is written as

$$\begin{aligned}\mathcal{L}(\pi_0, S_{\mathrm{ND}}, S_{\mathrm{D}}) &= \pi_0^{n_{\mathrm{ND}}}(1-\pi_0)^{n_{\mathrm{D}}} \prod_{i=\mathrm{ND},\mathrm{D}} \prod_{j=1}^{m+1} \bigl\{S_i(x_{j-1}) - S_i(x_j)\bigr\}^{d_{ij}} \\ &= \mathcal{L}(\pi_0)\mathcal{L}(S_{\mathrm{ND}}, S_{\mathrm{D}}),\end{aligned} \tag{3}$$

where π_0 is the prior probability of class D, and d_{ij} denotes the number of observations of class i in the semi-closed interval $(x_{j-1}, x_j]$. (We interpret $x_0 = -\infty$ and $x_{m+1} = \infty$ so that $S_i(x_0) = 1$ and $S_i(x_{m+1}) = 0$.) $\mathcal{L}(\pi_0)$ is maximized at $\hat{\pi}_0 = n_{\mathrm{D}}/(n_{\mathrm{ND}} + n_{\mathrm{D}})$ independent of $\mathcal{L}(S_{\mathrm{ND}}, S_{\mathrm{D}})$. It turns out that the maximizer of $\mathcal{L}(S_{\mathrm{ND}}, S_{\mathrm{D}})$ is the pair of empirical class conditional survival functions $(\hat{S}_{\mathrm{ND}}, \hat{S}_{\mathrm{D}})$. Hence the NPMLE of the ROC curve with no constraint is the empirical ROC curve

$$\hat{R}(p) = \hat{S}_{\mathrm{D}}\bigl(\hat{S}_{\mathrm{ND}}^{-1}(p)\bigr), \quad \text{for } p \in [0, 1], \tag{4}$$

or a plug-in estimator of (2). Note that only $\mathcal{L}(S_{\mathrm{ND}}, S_{\mathrm{D}})$ is needed to estimate the ROC curve.

2.2 Geometric programming formulation

The NPMLE of a *convex* ROC curve can be obtained by solving (3) after imposing appropriate constraints, and we show in this section that the resulting optimization problem is formulated as a geometric program (GP), a special class of convex optimization problems. What we want to solve is the following problem.

$$\begin{aligned}&\text{maximize} \quad \mathcal{L}(S_{\mathrm{ND}}, S_{\mathrm{D}}) = \prod_{i=\mathrm{ND},\mathrm{D}} \prod_{j=1}^{m+1} \bigl\{S_i(x_{j-1}) - S_i(x_j)\bigr\}^{d_{ij}} \\ &\text{subject to} \quad R(p) = S_{\mathrm{D}}\bigl(S_{\mathrm{ND}}^{-1}(p)\bigr) \text{ is convex in } p \in [0, 1].\end{aligned} \tag{5}$$

The solution to (5) is a pair of distributions that change their values only at the finite number of points $x_1, \dots, x_m$: if an estimated pair $(\tilde{S}_{\mathrm{ND}}^{\mathrm{con}}, \tilde{S}_{\mathrm{D}}^{\mathrm{con}})$ does not have such property, we could find an alternative solution satisfying the convexity constraint, whose likelihood is larger than that of $(\tilde{S}_{\mathrm{ND}}^{\mathrm{con}}, \tilde{S}_{\mathrm{D}}^{\mathrm{con}})$ (Kaplan and Meier 1958; Johansen 1978; Feltz and Dykstra 1985). Therefore we can fully specify convexity of the ROC curve in terms of the observed points and write problem (5) as

$$\begin{aligned} &\text{maximize} \quad \mathcal{L}(S_{\text{ND}}, S_{\text{D}}) = \prod_{i=\text{ND},\text{D}} \prod_{j=1}^{m+1} \bigl\{S_i(x_{j-1}) - S_i(x_j)\bigr\}^{d_{ij}} \\ &\text{subject to} \quad \frac{S_{\text{D}}(x_j) - S_{\text{D}}(x_{j-1})}{S_{\text{ND}}(x_j) - S_{\text{ND}}(x_{j-1})} \le \frac{S_{\text{D}}(x_{j+1}) - S_{\text{D}}(x_j)}{S_{\text{ND}}(x_{j+1}) - S_{\text{ND}}(x_j)}, \quad j = 1, \ldots, m. \end{aligned} \tag{6}$$

Note that S_{ND} and S_{D} are non-increasing in x. Now write

$$p_{ij} = S_i(x_j)/S_i(x_{j-1}), \quad i = \text{ND}, \text{D}, \; j = 1, \ldots, m.$$

so that $S_i(x_j) = \prod_{r=1}^{j} p_{ir}$, and introduce auxiliary variables $q_{ij} = 1 - p_{ij}$. Then (6) is written as:

$$\begin{aligned} &\text{maximize} \quad \mathcal{L}\bigl(\bigl\{(p_{ij}, q_{ij})\bigr\}\bigr) = \prod_{i=\text{ND},\text{D}} \prod_{j=1}^{m+1} q_{ij}^{d_{ij}} \prod_{r<j} p_{ir}^{d_{ij}} \\ &\text{subject to} \quad \left(\frac{p_{\text{ND},j}}{q_{\text{ND},j}}\right)\left(\frac{q_{\text{D},j}}{p_{\text{D},j}}\right)\left(\frac{q_{\text{ND},(j+1)}}{q_{\text{D},(j+1)}}\right) \le 1, \quad j = 1, \ldots, m, \\ &\qquad\qquad p_{ij} + q_{ij} = 1, \quad i = \text{ND}, \text{D}, \; j = 1, \ldots, m. \end{aligned} \tag{7}$$

If we further relax the equality constraints $p_{ij} + q_{ij} = 1$ with inequalities $p_{ij} + q_{ij} \le 1$, problem (7) becomes a GP:

$$\begin{aligned} &\text{maximize} \quad \mathcal{L}\bigl(\bigl\{(p_{ij}, q_{ij})\bigr\}\bigr) \\ &\text{subject to} \quad \left(\frac{p_{\text{ND},j}}{q_{\text{ND},j}}\right)\left(\frac{q_{\text{D},j}}{p_{\text{D},j}}\right)\left(\frac{q_{\text{ND},(j+1)}}{q_{\text{D},(j+1)}}\right) \le 1, \quad \text{for } j = 1, \ldots, m, \\ &\qquad\qquad p_{ij} + q_{ij} \le 1, \quad i = \text{ND}, \text{D}, \; j = 1, \ldots, m. \end{aligned} \tag{8}$$

A standard GP is not in general convex, but can be transformed to a convex form using a simple change of variables, e.g., $u_{ij} = \log p_{ij}$ and $v_{ij} = \log q_{ij}$ here. To see the equivalence of the relaxed GP formulation (8) to the original problem (7), observe that the quantity $p_{ij} + q_{ij}$ is monotone increasing in both p_{ij} and q_{ij} and the objective is increasing in these variables. We see that at the optimal point $\{(\bar{p}_{ij}, \bar{q}_{ij})\}$, the inequality constraints $p_{ij} + q_{ij} \le 1$ must be tight for all i and j. Otherwise there exist i' and j' with $\bar{p}_{i'j'} + \bar{q}_{i'j'} < 1$. Then $\{(\bar{p}_{ij}, \bar{q}_{ij})\}$ cannot be optimal since a point $\{(\hat{p}_{ij}, \hat{q}_{ij})\}$ with $\hat{p}_{ij} = \bar{p}_{ij}$ and $\hat{q}_{ij} = 1 - \bar{p}_{ij}$ for $i = \text{ND}, \text{D}$, $j = 1, \ldots, m$ is feasible for (8) and $\hat{q}_{i'j'} = 1 - \bar{p}_{i'j'} > \bar{q}_{i'j'}$. This results in

$$\mathcal{L}\bigl(\bigl\{(\hat{p}_{ij}, \hat{q}_{ij})\bigr\}\bigr) > \mathcal{L}\bigl(\bigl\{(\bar{p}_{ij}, \bar{q}_{ij})\bigr\}\bigr),$$

which is a contradiction. Therefore the two problems are equivalent.

2.3 NPMLE yields the ROCCH

The solution to the GP (8) is readily available as the ROCCH of the empirical ROC curve (4), without needing a numerical GP solver, e.g., `ggplab` (Mutapcic et al. 2006). To see this, write the full likelihood (3) in terms of class conditional densities, instead of survival functions, in two alternative factorizations:

$$\mathcal{L}(\pi_0, S_{\mathrm{ND}}, S_{\mathrm{D}}) = \mathcal{L}(\pi_0)\mathcal{L}(f_{\mathrm{ND}}, f_{\mathrm{D}}) = \pi_0^{n_{\mathrm{ND}}}(1-\pi_0)^{n_{\mathrm{D}}} \prod_{i=\mathrm{ND,D}} \prod_{j=1}^{n_i} f_i(x_{ij})^{d_{ij}} \tag{9}$$

$$= \prod_{j=1}^{m} \pi(x_j)^{d_{\mathrm{D},j}} \big(1-\pi(x_j)\big)^{d_{\mathrm{ND},j}} \prod_{j=1}^{m} f(x_j)^{d_{\mathrm{D},j}+d_{\mathrm{ND},j}} = \mathcal{L}(\pi)\mathcal{L}(f), \tag{10}$$

where $f_{\mathrm{ND}}(x)$ and $f_{\mathrm{D}}(x)$ are class conditional densities with $f_i(x_{ij}) = S_i(x_{j-1}) - S_i(x_j)$, $i = \mathrm{ND}, \mathrm{D}$;[2] $f(x) = (1-\pi_0)f_{\mathrm{ND}}(x) + \pi_0 f_{\mathrm{D}}(x)$ is the marginal density of score X; and $\pi(x)$ is the posterior class probability given $X = x$, so that $\pi_0 = \int \pi(x) f(x)dx$. Lloyd (2002) shows that the following two-step optimization procedure maximizes (9) (equivalently (10)) subject to the convexity constraint on the ROC curve being estimated, hence solves the GP (8).

Step 1: estimate $\pi(x)$ *nonparametrically* so that

$$\begin{array}{ll} \text{maximize} & \mathcal{L}(\pi) = \prod_{j=1}^{m} \pi(x_j)^{d_{\mathrm{D},j}} \big(1-\pi(x_j)\big)^{d_{\mathrm{ND},j}} \\ \text{subject to} & \pi(x) \text{ monotone nondecreasing.} \end{array}$$

Step 2: estimate f_{ND} and f_{D} so that

$$\begin{array}{ll} \text{maximize} & \mathcal{L}(f_{\mathrm{ND}}, f_{\mathrm{D}}) = \prod_{i=\mathrm{ND,D}} \prod_{j=1}^{n_i} f_i(x_{ij})^{d_{ij}} \\ \text{subject to} & f_{\mathrm{D}}(x)/f_{\mathrm{ND}}(x) \propto \pi(x)/\big(1-\pi(x)\big). \end{array}$$

The solution to Step 1 is specified by the discrete density $\hat{\pi}(x)$ that is the PAV isotonic regression of the observed proportion $d_{\mathrm{D},j}/(d_{\mathrm{ND},j}+d_{\mathrm{D},j})$ of the positive class at each $x = x_j$. The solution to Step 2 is given by

$$\hat{f}_{\mathrm{ND}}^{\mathrm{lloyd}}(x_j) = \begin{cases} (d_{\mathrm{ND},j}+d_{\mathrm{D},j})\mu/(n_{\mathrm{D}}\hat{\phi}(x_j) + n_{\mathrm{ND}}\mu), & \hat{\phi}(x_j) < \infty, \\ 0, & \hat{\phi}(x_j) = \infty, \end{cases} \tag{11}$$

and

$$\hat{f}_{\mathrm{D}}^{\mathrm{lloyd}}(x_j) = \begin{cases} (d_{\mathrm{ND},j}+d_{\mathrm{D},j})\hat{\phi}(x_j)/(n_{\mathrm{D}}\hat{\phi}(x_j) + n_{\mathrm{ND}}\mu), & \hat{\phi}(x_j) < \infty, \\ (d_{\mathrm{ND},j}+d_{\mathrm{D},j})/n_{\mathrm{D}}, & \hat{\phi}(x_j) = \infty, \end{cases} \tag{12}$$

where $\hat{\phi}(x) = \hat{\pi}(x)/(1-\hat{\pi}(x))$. $\hat{\phi}(x) = \infty$ if $\hat{\pi}(x) = 1$. μ is chosen so that both $\hat{f}_{\mathrm{ND}}^{\mathrm{lloyd}}(x_j)$ and $\hat{f}_{\mathrm{D}}^{\mathrm{lloyd}}(x_j)$ sum to one. This solution yields an estimate of the class probability $\hat{\pi}_0 = n_{\mathrm{D}}/(n_{\mathrm{ND}} + n_{\mathrm{D}})$. The NPMLE of the convex ROC curve is obtained by reconstruct-

[2] The existence of the class conditional density function and writing it in this form is supported by that S_i changes its value only at the points $x_1, \ldots, x_m$; see Sect. 2.2.

ing the class conditional survival functions $\tilde{S}_{\mathrm{D}}^{\mathrm{lloyd}}(x_j) = \sum_{l>j} \hat{f}_{\mathrm{D}}^{\mathrm{lloyd}}(x_l)$ and $\tilde{S}_{\mathrm{ND}}^{\mathrm{lloyd}}(x_j) = \sum_{l>j} \hat{f}_{\mathrm{ND}}^{\mathrm{lloyd}}(x_l)$, and plotting $\tilde{S}_{\mathrm{D}}^{\mathrm{lloyd}}(x_j)$ against $\tilde{S}_{\mathrm{ND}}^{\mathrm{lloyd}}(x_j)$.

To see why this estimate coincides with the ROCCH, it suffices to recognize that $\hat{\pi}(x)$, the PAV isotonic regression of the proportion of the positive class at a given x, is essentially the classifier score calibrated using the same regression method (Zadrozny and Elkan 2002). The estimated ROC curve is that determined by the calibrated scores. This is precisely the ROCCH, because of the equivalence between the ROCCH and the PAV regression-based calibration as discussed in Sect. 1 (for more details, see Fawcett and Niculescu-Mizil 2007). This connection between the NPMLE and the ROCCH has not been known previously.

The GP formulation (8) can be used to impose a wider class of constraints to the NPMLE problem, e.g., ordering of several convex ROC curves that establishes superiority of a classifier to another; for various order constraints in GP, see Lim et al. (2009). More importantly in this paper, the GP (8) provides a crucial insight leading to the results of the next section.

3 Conditional NPMLE that yields the ROCCH

Surprisingly, even if we condition that each FPR estimate is equal to the corresponding empirical FPR, the resulting NPMLE of the convex ROC curve still coincides with the ROCCH. To be specific, assume that $S_{\mathrm{ND}} := \hat{S}_{\mathrm{ND}}$. Let $\{\nu_1, \dots, \nu_l\}$, $\nu_1 < \cdots < \nu_l$, be a subset of $\{1, \dots, m\}$ such that each $\hat{S}_{\mathrm{ND}}(x_{\nu_j})$ is unique, i.e., $\hat{S}_{\mathrm{ND}}(x_{\nu_{j-1}}) \neq \hat{S}_{\mathrm{ND}}(x_{\nu_j})$ for any j. By the assumption, we set $p_{\mathrm{ND},j}$ as its empirical estimate

$$\hat{p}_{\mathrm{ND},j} = \hat{S}_{\mathrm{ND}}(x_{\nu_j}) / \hat{S}_{\mathrm{ND}}(x_{\nu_{j-1}}) \tag{13}$$

for $j = 1, \dots, l$. (We interpret $x_{\nu_0} = -\infty$ and $x_{\nu_{l+1}} = \infty$ so that $S_i(x_{\nu_0}) = 1$ and $S_i(x_{\nu_{l+1}}) = 0$, $i = \mathrm{ND}, \mathrm{D}$.) The conditional NPMLE is then formulated as follows.

$$\begin{aligned} &\text{maximize} && \mathcal{L}(\{p_{\mathrm{D},j}\}) = \prod_{j=1}^{l+1} (1 - p_{\mathrm{D},j})^{d_{\mathrm{D},j}} \prod_{r<j} p_{\mathrm{D},r}^{d_{\mathrm{D},j}} \\ &\text{subject to} && \left(\frac{\hat{p}_{\mathrm{ND},j}}{1 - \hat{p}_{\mathrm{ND},j}}\right)\left(\frac{1 - p_{\mathrm{D},j}}{p_{\mathrm{D},j}}\right)\left(\frac{1 - \hat{p}_{\mathrm{ND},(j+1)}}{1 - p_{\mathrm{D},(j+1)}}\right) \leq 1, \quad \text{for } j = 1, \dots, l, \end{aligned} \tag{14}$$

where variables are $\{p_{\mathrm{D},j}\}$, $p_{\mathrm{D},j} = S_{\mathrm{D}}(x_{\nu_j}) / S_{\mathrm{D}}(x_{\nu_{j-1}})$. Note that each $d_{\mathrm{D},j}$ is appropriately redefined to be the number of observations of class D in the semi-closed interval $(x_{\nu_{j-1}}, x_{\nu_j}]$ of scores. We refer to this problem as conditional NPMLE. Note that (14) can also be rewritten as a GP in a similar fashion to the unconditional NPMLE (7).

That the ROCCH, or the LCM of the empirical ROC curve, is the conditional NPMLE of the convex ROC curve can be summarized by the following theorem.

Theorem 1 *Let $\tilde{S}_{\mathrm{D}}^{\mathrm{lcm}}$ be the class conditional survival function such that the curve $(\hat{S}_{\mathrm{ND}}, \tilde{S}_{\mathrm{D}}^{\mathrm{lcm}})$ is the LCM of the empirical ROC curve $(\hat{S}_{\mathrm{ND}}, \hat{S}_{\mathrm{D}})$. Then, $\tilde{S}_{\mathrm{D}}^{\mathrm{lcm}}$ solves the conditional NPMLE (14). More precisely, $\tilde{p}_{\mathrm{D},j} = \tilde{S}_{\mathrm{D}}^{\mathrm{lcm}}(x_{\nu_j}) / \tilde{S}_{\mathrm{D}}^{\mathrm{lcm}}(x_{\nu_{j-1}})$ solves (14).*

Proof We consider an iterative (coordinate ascent) procedure to solve (14), which iteratively updates $p_{\mathrm{D},k}$ by maximizing (14) with respect to $p_{\mathrm{D},k}$ for $k = 1, \dots, l$. In updating $p_{\mathrm{D},k}$, all other $p_{\mathrm{D},j}$ with $j \neq k$ are held fixed at their current estimates. Since the problem (14) is equivalent to a GP, which can be converted to a convex problem, and the auxiliary variables

$q_{\mathrm{D},j}$, introduced to construct the GP, satisfy $q_{\mathrm{D},j} = 1 - p_{\mathrm{D},j}$ for all j (see Sect. 2.2), the suggested coordinate ascent procedure will solve the problem. It is easy to see that for each step of the iterative procedure solves the following subproblem.

$$\begin{aligned} &\text{maximize} \quad n_{\mathrm{D},k} \log p_{\mathrm{D},k} + d_{\mathrm{D},k} \log(1 - p_{\mathrm{D},k}) \\ &\text{subject to} \quad L_k \le p_{\mathrm{D},k} \le U_k, \end{aligned} \tag{15}$$

where $n_{\mathrm{D},j} = \sum_{r=j+1}^{l+1} d_{\mathrm{D},r}$ denotes the number of observations of class D whose scores are greater than x_{ν_j}; L_k and U_k are bounds determined by the other coordinates $p_{\mathrm{D},j}$, $j \ne k$.

By construction, it suffices to show that $\{\tilde{p}_{\mathrm{D},j}\}$ is a fixed point of the iterative procedure. For each k, we fix all the coordinates except for $p_{\mathrm{D},k}$ at $\tilde{p}_{\mathrm{D},j}$, i.e., $p_{\mathrm{D},j} = \tilde{p}_{\mathrm{D},j}$, $j \ne k$. Then,

$$L_k = \frac{\{\hat{p}_{\mathrm{ND},k}(1 - \hat{p}_{\mathrm{ND},(k+1)})\}/\{(1 - \hat{p}_{\mathrm{ND},k})(1 - \tilde{p}_{\mathrm{D},(k+1)})\}}{1 + \{\hat{p}_{\mathrm{ND},k}(1 - \hat{p}_{\mathrm{ND},(k+1)})\}/\{(1 - \hat{p}_{\mathrm{ND},k})(1 - \tilde{p}_{\mathrm{D},(k+1)})\}},$$

for $k = 1, \ldots, l$, and

$$U_k = 1 - \frac{\tilde{p}_{\mathrm{ND},(k-1)}(1 - \tilde{p}_{\mathrm{ND},k})(1 - \tilde{p}_{\mathrm{D},(k-1)})}{\tilde{p}_{\mathrm{D},(k-1)}(1 - \tilde{p}_{\mathrm{ND},(k-1)})}$$

for $k = 2, \ldots, l$. For $k = 1$, we set $U_1 = 1$.

Now consider

$$\hat{p}_{\mathrm{D},j} = \hat{S}_{\mathrm{D}}(x_{\nu_j})/\hat{S}_{\mathrm{D}}(x_{\nu_{j-1}}), \quad j = 1, \ldots, l.$$

Together with $\{\hat{p}_{\mathrm{ND},j}\}$ defined in (13), $\{\hat{p}_{\mathrm{D},j}\}$ constitutes the empirical ROC, which solves the unconstrained version of the unconditional NPMLE (7). Therefore $\{\hat{p}_{\mathrm{D},j}\}$ solves the unconstrained version of the conditional NPMLE (14). It follows that $\hat{p}_{\mathrm{D},k}$ maximizes the objective of (15) provided the constraint was removed. Let $\hat{p}_{\mathrm{D},k}^{\mathrm{local}}$ denote the (constrained) solution to (15). Showing that $\hat{p}_{\mathrm{D},k}^{\mathrm{local}} = \tilde{p}_{\mathrm{D},k}$ completes the proof.

The following property of $\tilde{p}_{\mathrm{D},k}$ locally characterizes not only the LCM, but also the empirical ROC curve in the neighborhood of $[x_{\nu_{k-1}}, x_{\nu_k}]$, leading to identification of the solution $\hat{p}_{\mathrm{D},k}^{\mathrm{local}}$. Observe that $\tilde{p}_{\mathrm{D},k} < U_k$ if and only if the inequality in the convexity constraint in (5) is strict for $j = k - 1$. In other words, the LCM changes its slope at $(\hat{S}_{\mathrm{ND}}(x_{\nu_{k-1}}), \tilde{S}_{\mathrm{D}}^{\mathrm{lcm}}(x_{\nu_{k-1}}))$. Change of slope of the LCM occurs if and only if it touches the empirical ROC curve, hence we have $\tilde{S}_{\mathrm{D}}^{\mathrm{lcm}}(x_{\nu_{k-1}}) = \hat{S}_{\mathrm{D}}(x_{\nu_{k-1}})$ (note that in general $\tilde{S}_{\mathrm{D}}^{\mathrm{lcm}}(x) \ge \hat{S}_{\mathrm{D}}(x)$). Similarly, $L_k < \hat{p}_{\mathrm{D},k}^{\mathrm{lcm}}$ if and only if $\tilde{S}_{\mathrm{D}}^{\mathrm{lcm}}(x_{\nu_k}) = \hat{S}_{\mathrm{D}}(x_{\nu_k})$. Since the LCM is convex by construction, $\tilde{p}_{\mathrm{D},k}$ always satisfies $L_k \le \tilde{p}_{\mathrm{D},k} \le U_k$. Depending on the tightness of these bounds, there are four cases to consider:

1. $L_k < \tilde{p}_{\mathrm{D},k} < U_k$: From the observation above, $\tilde{S}_{\mathrm{D}}^{\mathrm{lcm}}(x_{\nu_{k-1}}) = \hat{S}_{\mathrm{D}}(x_{\nu_{k-1}})$ and $\tilde{S}_{\mathrm{D}}^{\mathrm{lcm}}(x_{\nu_k}) = \hat{S}_{\mathrm{D}}(x_{\nu_k})$. Then,

$$\tilde{p}_{\mathrm{D},k} = \tilde{S}_{\mathrm{D}}^{\mathrm{lcm}}(x_{\nu_k})/\tilde{S}_{\mathrm{D}}^{\mathrm{lcm}}(x_{\nu_{k-1}}) = \hat{S}_{\mathrm{D}}(x_{\nu_k})/\hat{S}_{\mathrm{D}}(x_{\nu_{k-1}}) = \hat{p}_{\mathrm{D},k},$$

i.e., $L_k < \hat{p}_{\mathrm{D},k} < U_k$. Since $\hat{p}_{\mathrm{D},k}$ is the unconstrained maximizer of the objective of (15), which is convex in $p_{\mathrm{D},k}$, $\hat{p}_{\mathrm{D},k}$ also solves the constrained problem (15). Therefore $\hat{p}_{\mathrm{D},k}^{\mathrm{local}} = \hat{p}_{\mathrm{D},k} = \tilde{p}_{\mathrm{D},k}$.

2. $L_k = \tilde{p}_{\mathrm{D},k} < U_j$: We have

$$\tilde{S}_{\mathrm{D}}^{\mathrm{lcm}}(x_{v_{k-1}}) = \hat{S}_{\mathrm{D}}(x_{v_{k-1}}), \quad \text{or} \quad \prod_{j=1}^{k-1} \tilde{p}_{\mathrm{D},j} = \prod_{j=1}^{k-1} \hat{p}_{\mathrm{D},j}, \tag{16}$$

$$\tilde{S}_{\mathrm{D}}^{\mathrm{lcm}}(x_{v_k}) > \hat{S}_{\mathrm{D}}(x_{v_k}), \quad \text{or} \quad \prod_{j=1}^{k} \tilde{p}_{\mathrm{D},j} > \prod_{j=1}^{k} \hat{p}_{\mathrm{D},j}. \tag{17}$$

   Dividing (17) by (16) we obtain

$$\hat{p}_{\mathrm{D},k} < \tilde{p}_{\mathrm{D},k} = L_k, \tag{18}$$

   i.e., the unconstrained maximizer $\hat{p}_{\mathrm{D},k}$ of (15) is less than L_k. Combined with the convexity of the objective, this implies that the constrained maximizer of (15) satisfies $\hat{p}_{\mathrm{D},k}^{\mathrm{local}} = L_k = \tilde{p}_{\mathrm{D},k}$.
3. $L_k < \tilde{p}_{\mathrm{D},k} = U_j$: This case is essentially the same as case 2, with L_k replaced by U_k and the inequality in (18) reversed.
4. $L_k = \tilde{p}_{\mathrm{D},k} = U_k$: Since $L_k = U_k$, the solution to the constrained maximizer of (15) is $\hat{p}_{\mathrm{D},k}^{\mathrm{local}} = L_k = U_k = \tilde{p}_{\mathrm{D},k}$.

Therefore we have $\hat{p}_{\mathrm{D},k}^{\mathrm{local}} = \tilde{p}_{\mathrm{D},k}$ in all four cases. □

Although both the conditional and the unconditional NPMLEs result in the ROCCH as the estimated ROC curve, the two methods produce distinct estimates of the TPR and the FPR. In general, unconditional NPMLE gives smoother estimates of these since it gives distinct scores to the samples. To understand this, we show in Table 1 estimated quantities including TPRs and FPRs using these two NPMLE methods from the example presented in Fawcett and Niculescu-Mizil (2007). The first two columns represent the observation, where the score is sorted in decreasing order. Class label 1 corresponds to the positive class (D), and 0 to the negative class (ND). The third and the fourth columns are the empirical FPRs and TPRs, so that $(\hat{S}_{\mathrm{ND}}, \hat{S}_{\mathrm{D}})$ constitutes the empirical ROC curve. The fifth column consists of numerical solutions of the GP (14) given the empirical FPRs, so that $(\hat{S}_{\mathrm{ND}}, \tilde{S}_{\mathrm{D}}^{\mathrm{lcm}})$ constitutes the conditional NPMLE. Columns 6 through 11 are for the unconditional NPMLE computed using Lloyd's method discussed in Sect. 2.3. In particular, the last two columns $\tilde{S}_{\mathrm{ND}}^{\mathrm{lloyd}}$ and $\tilde{S}_{\mathrm{D}}^{\mathrm{lloyd}}$ together make up the unconditional NPMLE of the convex ROC curve. The boldfaced indicates that both NPMLEs coincide with the ROCCH and meet the empirical ROC curve at the same points. These points are the vertices of the ROCCH. The conditional NPMLE estimates the TPRs effectively only at these vertex points, whereas the unconditional NPMLE estimates them (and the FPRs) in between.

Finally, it is worth note that from the (anti-)symmetry in the formulation (6) and the proof of Theorem 1 we can obtain the ROCCH by fixing the TPR estimate:

Corollary 1 *Let $\tilde{S}_{\mathrm{ND}}^{\mathrm{lcm}}$ be the class conditional survival function such that the curve $(\tilde{S}_{\mathrm{ND}}^{\mathrm{lcm}}, \hat{S}_{\mathrm{D}})$ is the least* concave[3] *majorant of the empirical ROC curve $(\hat{S}_{\mathrm{ND}}, \hat{S}_{\mathrm{D}})$ seen from the TPR axis. Then, $\tilde{S}_{\mathrm{ND}}^{\mathrm{lcm}}$ solves the conditional NPMLE (14) with the roles of ND and D switched.*

[3] Used as the opposite to the notion of "convex" as described in footnote 1.

Table 1 An illustration of the unconditional and the conditional NPMLEs

1	2	3	4	5	6	7	8	9	10	11
Score	Class	$\hat{S}_{\mathrm{ND}}$	$\hat{S}_{\mathrm{D}}$	$\tilde{S}_{\mathrm{D}}^{\mathrm{lcm}}$	$\hat{\pi}$	$\hat{\phi}$	$\hat{f}_{\mathrm{ND}}^{\mathrm{lloyd}}$	$\hat{f}_{\mathrm{D}}^{\mathrm{lloyd}}$	$\tilde{S}_{\mathrm{ND}}^{\mathrm{lloyd}}$	$\tilde{S}_{\mathrm{D}}^{\mathrm{lloyd}}$
0.9	1	0	0	0.2222	1	∞	0	0.1111	0	0
0.8	1	0	0.1111	0.2222	1	∞	0	0.1111	0	0.1111
0.7	0	**0**	**0.2222**	**0.2222**	0.75	3	0.0417	0.0833	**0**	**0.2222**
0.6	1	0.1667	0.2222	0.5556	0.75	3	0.0417	0.0833	0.0417	0.3056
0.55	1	0.1667	0.3333	0.5556	0.75	3	0.0417	0.0833	0.0833	0.3889
0.5	1	0.1667	0.4444	0.5556	0.75	3	0.0417	0.0833	0.125	0.4722
0.45	0	**0.1667**	**0.5556**	**0.5556**	0.6667	2	0.0556	0.0741	**0.1667**	**0.5556**
0.4	1	0.3333	0.5556	0.7778	0.6667	2	0.0556	0.0741	0.2222	0.6296
0.35	1	0.3333	0.6667	0.7778	0.6667	2	0.0556	0.0741	0.2778	0.7037
0.3	0	**0.3333**	**0.7778**	**0.7778**	0.5	1	0.0833	0.0556	**0.3333**	**0.7778**
0.27	1	0.5	0.7778	0.8889	0.5	1	0.0833	0.0556	0.4167	0.8333
0.2	0	**0.5**	**0.8889**	**0.8889**	0.3333	0.5	0.1111	0.0370	**0.5**	**0.8889**
0.18	0	0.6667	0.8889	0.9444	0.3333	0.5	0.1111	0.0370	0.6111	0.9259
0.1	1	0.8333	0.8889	1	0.3333	0.5	0.1111	0.0370	0.7222	0.9630
0.02	0	**0.8333**	**1**	**1**	0	0	0.1667	0	**0.8333**	**1**

4 Conditional bootstrap of the ROCCH

That the conditional NPMLE of a convex ROC curve coincides with the ROCCH of the empirical ROC curve suggests an useful bootstrap procedure to estimate the variance of the ROCCH (see, e.g., Macskassy et al. 2005). This *conditional bootstrap* procedure, which samples separately from the positive and the negative groups, allows us to evaluate the variance component contributed by each of the groups being compared (Hinkley 1988; Tibshirani and Knight 1999). Decomposing the variance components is advantageous because each term constitutes an achievable minimum *total* variance of the ROCCH estimate when the size of the corresponding group increases. Therefore, this procedure also provides a simple means to compute the total variance of the ROCCH when the sample is imbalanced (Mladenic and Grobelnik 1999). The pointwise confidence limit for a convex ROC curve $R(p) = S_{\mathrm{D}}(S_{\mathrm{ND}}^{-1}(p))$ relies on the variance of its conditional NPMLE $\tilde{\mathbf{R}}(p) = \tilde{\mathbf{S}}_{\mathrm{D}}(\hat{\mathbf{S}}_{\mathrm{ND}}^{-1}(p))$, or the ROCCH. (We use boldface letters to emphasize that the corresponding quantities are random. Normal-faced letters are their realizations.) From the law of total variance, the variance of the ROCCH can be decomposed as

$$\mathrm{Var}\big[\tilde{\mathbf{S}}_{\mathrm{D}}\big(\hat{\mathbf{S}}_{\mathrm{ND}}^{-1}(p)\big)\big] = E\big[\mathrm{Var}\big[\tilde{\mathbf{S}}_{\mathrm{D}}\big(\hat{\mathbf{S}}_{\mathrm{ND}}^{-1}(p)\big)|\hat{\mathbf{S}}_{\mathrm{ND}}\big]\big] + \mathrm{Var}\big[E\big[\tilde{\mathbf{S}}_{\mathrm{D}}\big(\hat{\mathbf{S}}_{\mathrm{ND}}^{-1}(p)\big)|\hat{\mathbf{S}}_{\mathrm{ND}}\big]\big], \qquad (19)$$

where the first and the second terms indicate the sampling variability from the negative (ND) and the positive (D) groups, respectively. The expectations in (19) can be approximated using a mode (or mode-type) approximation as

$$\begin{aligned} E\big[\mathrm{Var}\big[\tilde{\mathbf{S}}_{\mathrm{D}}\big(\hat{\mathbf{S}}_{\mathrm{ND}}^{-1}(p)\big)|\hat{\mathbf{S}}_{\mathrm{ND}}\big]\big] &\approx \mathrm{Var}\big[\tilde{\mathbf{S}}_{\mathrm{D}}\big(\mathbf{S}_{\mathrm{ND}}^{-1}(p)\big)|\mathbf{S}_{\mathrm{ND}}\big]\big|_{\mathbf{S}_{\mathrm{ND}}=\hat{S}_{\mathrm{ND}}} \\ &= \mathrm{Var}\big[\tilde{\mathbf{S}}_{\mathrm{D}}\big(\hat{S}_{\mathrm{ND}}^{-1}(p)\big)\big] \end{aligned} \qquad (20)$$

and

$$E\big[\tilde{\mathbf{S}}_{\mathrm{D}}\big(\hat{\mathbf{S}}_{\mathrm{ND}}^{-1}(p)\big)|\hat{\mathbf{S}}_{\mathrm{ND}}\big] \approx \mathbf{S}_{\mathrm{D}}\big(\tilde{\mathbf{S}}_{\mathrm{ND}}^{-1}(p)\big)\big|_{\mathbf{S}_{\mathrm{D}}=\hat{S}_{\mathrm{D}}} = \hat{S}_{\mathrm{D}}\big(\tilde{\mathbf{S}}_{\mathrm{ND}}^{-1}(p)\big), \tag{21}$$

where $\hat{S}_{\mathrm{D}}$ and $\hat{S}_{\mathrm{ND}}$ (normal-faced) are the observed TPR and FPR. Now it is seen that

$$\mathrm{Var}\big[\tilde{\mathbf{S}}_{\mathrm{D}}\big(\tilde{\mathbf{S}}_{\mathrm{ND}}^{-1}(p)\big)\big] \approx \mathrm{Var}\big[\tilde{\mathbf{S}}_{\mathrm{D}}\big(\hat{S}_{\mathrm{ND}}^{-1}(p)\big)\big] + \mathrm{Var}\big[\hat{S}_{\mathrm{D}}\big(\tilde{\mathbf{S}}_{\mathrm{ND}}^{-1}(p)\big)\big], \tag{22}$$

a separation of the contribution to the variance from the positive and the negative groups, respectively. Observe that the first term (resp. the second term) of the right-hand side becomes the achievable minimum total variance (the left-hand side) by increasing n_{D} (resp. n_{ND}); the second term (resp. the first term) degenerates to zero as n_{D} (resp. n_{ND}) increases.

The first ("positive") term is computed as follows.

Algorithm 1: Conditional bootstrap method for variance component decomposition

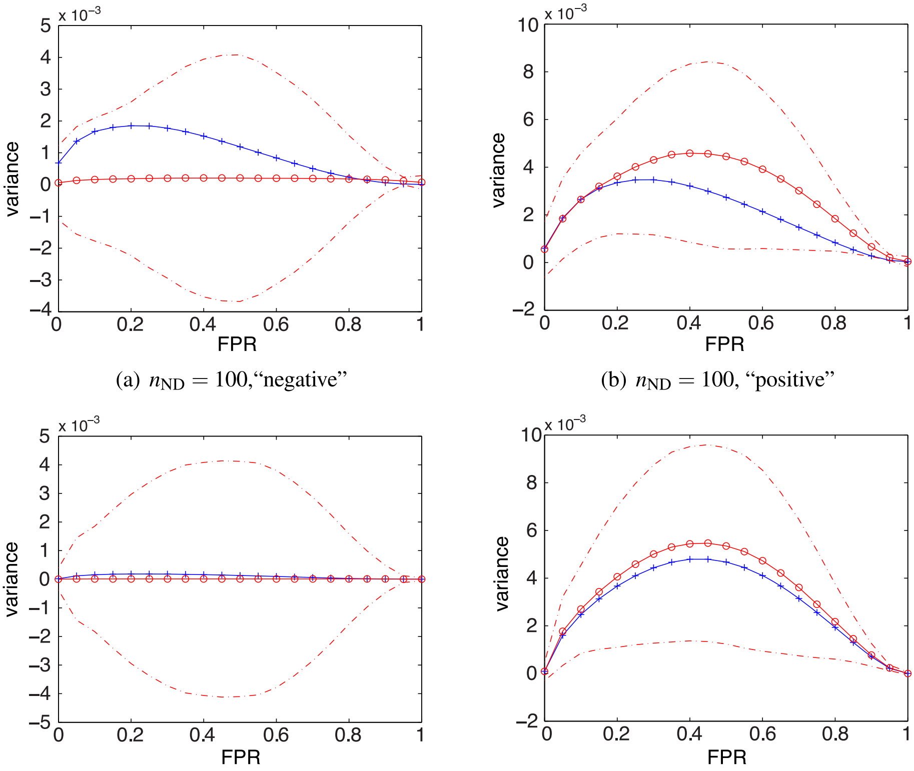
```
input : Negative samples ${x_{ND,1}, ..., x_{ND,n_{ND}}}$; Positive samples ${x_{D,1}, ..., x_{D,n_D}}$
        Bootstrap sample size $B$
output: Estimate of the variance component $\mathrm{Var}[\tilde{\mathbf{S}}_{\mathrm{D}}(\hat{S}_{\mathrm{ND}}^{-1}(p))]$ of the ROCCH due to
        the positive group
1 begin
2     Fix the negative group;
3     for $b \leftarrow 1$ to $B$ do
4         Bootstrap from the positive group only;
5         Estimate the ROCCH $\tilde{S}_{\mathrm{D}}^{(b)}(\hat{S}_{\mathrm{ND}}^{-1}(p))$ of the bootstrapped samples;
6     end
7     Output the pointwise sample variance of $\{\tilde{S}_{\mathrm{D}}^{(b)}(\hat{S}_{\mathrm{ND}}^{-1}(p))\}_{b=1}^{B}$;
8 end
```

Note that Theorem 1 takes action in line 5. For the second ("negative") term, switch the positive and the negative groups in the above procedure. Note that this separate evaluation is not possible in a naive bootstrap, that resamples the whole $n_{\mathrm{D}} + n_{\mathrm{ND}}$ observations. When the sample is imbalanced, e.g., $n_{\mathrm{D}} \ll n_{\mathrm{ND}}$, then

$$\mathrm{Var}\big[\hat{S}_{\mathrm{D}}\big(\tilde{\mathbf{S}}_{\mathrm{ND}}^{-1}(p)\big)\big] \approx 0 \quad \text{and} \quad \mathrm{Var}\big[\tilde{\mathbf{S}}_{\mathrm{D}}\big(\tilde{\mathbf{S}}_{\mathrm{ND}}^{-1}(p)\big)\big] \approx \mathrm{Var}\big[\tilde{\mathbf{S}}_{\mathrm{D}}\big(\hat{S}_{\mathrm{ND}}^{-1}(p)\big)\big],$$

so that only the "positive" conditional bootstrap suffices.

We conducted a simple numerical study to demonstrate how the proposed conditional bootstrap procedure approximates the variance components in (19) and how these components vary as the sample gets imbalanced. We set $n_{\mathrm{D}} = 50$ and varied $n_{\mathrm{ND}} = 50$, 100, 200, 300, and 1000. The scores of the negative group were distributed normally with mean 0 and variance 1, and the scores of the positive group are distributed normally with mean 0.5 and variance 1. For each choice of n_{ND}, we generated $B = 500$ data sets and applied the conditional bootstrap to estimate the "positive" and the "negative" variance terms. We compared them with their true values in (19). The results are shown in Fig. 1 for $n_{\mathrm{ND}} = 100$ and 1000. Note that the bootstrap estimates of both variance terms are very close to the true values. In particular, the "negative" variance estimate almost vanishes at $n_{\mathrm{ND}} = 1000$.



(a) $n_{\mathrm{ND}} = 100$,“negative”

(b) $n_{\mathrm{ND}} = 100$, “positive”

(c) $n_{\mathrm{ND}} = 1000$, “negative”

(d) $n_{\mathrm{ND}} = 1000$, “positive”

Fig. 1 Illustration of conditional bootstrap for $n_{\mathrm{D}} = 50$. In the figure, the *circle* indicates the true variance; the *cross* indicates bootstrap estimate of the variance; the *dash-dot* indicates the 5 % and 95 % bootstrap confidence limits of the variance, obtained from 500 bootstrap samples

5 Conclusion

In this paper we interpreted the ROC convex hull, which has been known as an efficient tool to account for the class-dependent misclassification cost in designing a classifier, from a maximum likelihood estimation perspective. We provided two nonparametric maximum likelihood formulations subject to the convexity constraint on the ROC curve and showed that the ROCCH is derived as the solution to both NPMLE problems. In particular the conditional NPMLE interpretation of the ROCCH enables standard machinery, such as the bootstrap, to assess uncertainties in the ROCCH. The proposed conditional bootstrap method can estimate the finite-sample variabilities of the ROCCH arising from the positive and the negative class separately, and allow us to find the achievable confidence limit of the ROCCH efficiently for imbalanced samples.

Acknowledgements J.L.’s research was supported by a Basic Science Research Program through the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education, Science, and Technology (No. 2010-0011448). J.-H.W. was partially supported by the US National Institutes of Health (NIH) MERIT Award R37EB02784. He was also supported by Dean’s Postdoctoral Fellowship of Stanford Medical School.

References

Campbell, G., & Ratnaparkhi, M. V. (1993). An application of Lomax distributions in receiver operating characteristic (ROC) curve analysis. *Communications in Statistics. Theory and Methods*, *22*(6), 1681–1697.

Dorfman, D. D., Berbaum, K. S., Metz, C. E., Lenth, R. V., Hanley, J. A., & Dagga, H. A. (1997). Proper receiver operating characteristic analysis: the bigamma model. *Academic Radiology*, *4*(2), 138–149.

Fawcett, T., & Niculescu-Mizil, A. (2007). PAV and the ROC convex hull. *Machine Learning*, *68*(1), 97–106.

Feltz, C. J., & Dykstra, R. L. (1985). Maximum likelihood estimation of the survival functions of n stochastically ordered random variables. *Journal of the American Statistical Association*, *80*(392), 1012–1019.

Flach, P. A., & Wu, S. (2005). Repairing concavities in ROC curves. In L. P. Kaelbling & A. Saffiotti (Eds.), *IJCAI* (pp. 702–707). Denver: Professional Book Center. ISBN 0938075934.

Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, *77*(1), 103–123.

Hinkley, D. V. (1988). Bootstrap methods. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *50*(3), 321–337.

Johansen, S. (1978). The product limit estimator as maximum likelihood estimator. *Scandinavian Journal of Statistics*, 195–199.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, *53*(282), 457–481.

Kim, S., Magnani, A., Samar, S., Boyd, S., & Lim, J. (2006). Pareto optimal linear classification. In *Proceedings of the 23rd international conference on machine learning* (pp. 473–480), Pittsburgh, Pennsylvania. New York: ACM.

Lim, J., & Pyun, K. (2009). Cost-effective hidden Markov model-based image segmentation. *IEEE Signal Processing Letters*, *16*(3), 172–175.

Lim, J., Kim, S. J., & Wang, X. (2009). Estimating stochastically ordered survival functions via geometric programming. *Journal of Computational and Graphical Statistics*, *18*(4), 978–994.

Lloyd, C. J. (2002). Estimation of a convex ROC curve. *Statistics & Probability Letters*, *59*(1), 99–111.

Macskassy, S., Provost, F., & Rosset, S. (2005). ROC confidence bands: an empirical evaluation. In *Proceedings of the 22nd international conference on machine learning* (pp. 537–544). New York: ACM.

Metz, C. E., & Pan, X. (1999). “Proper” binormal ROC curves: theory and maximum likelihood estimation. *Journal of Mathematical Psychology*, *43*, 1–33.

Metz, C. E., Herman, B. A., & Shen, J. (1998). Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Statistics in Medicine*, *17*(9), 1033–1053.

Mladenic, D., & Grobelnik, M. (1999). Feature selection for unbalanced class distribution and naive Bayes. In *Proceedings of the 16th international conference on machine learning*.

Mutapcic, A., Koh, K., Kim, S.-J., & Boyd, S. (2006). GGPLAB: a simple Matlab toolbox for geometric programming, version 1.00. http://stanford.edu/~boyd/ggplab, May 2006.

Pan, X., & Metz, C. E. (1997). The “proper” binormal model: parametric receiver operating characteristic curve estimation with degenerate data. *Academic Radiology*, *4*(5), 380–389.

Provost, F., & Fawcett, T. (2001). Robust classification for imprecise environments. *Machine Learning*, *42*(3), 203–231.

Tibshirani, R., & Knight, K. (1999). The covariance inflation criterion for adaptive model selection. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *61*(3), 529–546.

Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 694–699), Edmonton, Alberta, Canada. New York: ACM.