Statistica Sinica **29** (2019), 1395-1418
doi:https://doi.org/10.5705/ss.202014.0057

ON FEATURE ENSEMBLE OPTIMIZING THE SENSITIVITY AND PARTIAL ROC CURVE

Zheng Zhang[1], Ying Lu[2] and Lu Tian[2]

[1]*Peking University and* [2]*Stanford University*

Abstract: We consider a setting in which we construct a binary classifier from a panel of features in order to optimize either the sensitivity at a fixed specificity level or the area under the partial receiver operating characteristic (ROC) curve. To this end, we propose an efficient iterative numerical algorithm to solve a simple constrained optimization problem that mimics the original target. We also present the associated asymptotic statistical inference procedures, including the construction of the credible intervals for the realized sensitivity/specificity or the area under the partial ROC curve of the estimated risk scores. We apply the method to simulated data sets and show that the proposed method outperforms the classifiers based on the generic logistic regression, without considering the specific criterion we want to optimize. We also apply the new proposed method to two real-data examples.

Key words and phrases: Feature ensemble, ROC curve, sensitivity, specificity.

1. Introduction

Researchers often need to predict a binary outcome based on a collection of features. For example, the credit rating model identifies the credit card defaulters. The rating system is essentially a classification tool that signals the possible future status of individuals of interest. The rating score for each individual is calculated using features that characterize the borrower and the debt and aims to reflect the risk of default. Similarly, in a different setting, the Framingham risk score identifies people susceptible to a future cardiovascular attack based on baseline risk factors, including age, gender, blood pressure, and cholesterol level, among others (Wilson et al. (1998)). In general, there are two important tasks related to this type of application: (1) the development of a scoring system that measures the probability of an event occurring or the rank thereof; and (2) the evaluation of the effectiveness of the scoring system. Oftentimes, these two tasks are accomplished in separate stages. For example, at the first stage, we may fit a logistic regression model to associate a linear combination of features with the

binary outcome. At the second stage, the resulting scoring system, that is, the estimated linear combination (or a transformation thereof), is then evaluated using selected criteria. A scoring system with good discriminatory power groups individuals with similar risks together and assigns a higher score to individuals with higher risk. Quantitative quality measures for such scoring systems include the misclassification error, Brier score, and, importantly, sensitivity and specificity (Brier (1950)). Specifically, let Y and S be the binary response and risk score (or the rating), respectively. The sensitivity and specificity associated with a cutoff value d_0 are defined as $P(S \geq d_0|Y = 1)$ and $P(S \leq d_0|Y = 0)$, respectively. In practice, the cutoff value d_0 is often selected to guarantee a given specificity level, for example, $P(S \leq d_0|Y = 0) = \pi_0$. Then the corresponding sensitivity is used to measure the predictiveness of the risk score. In particular, when the event of interest is relatively rare, we often need to set a high specificity level, say $\pi_0 = 0.95$, in order to reduce the false positive rate of the classification tool. Furthermore, rather than a single specificity level, we might need to consider a range of possible levels, for exmaple, $[\pi_L, \pi_U]$. In this case, we may want to use the "average sensitivity" for all specificities within the interval to measure the quality of the score. In fact, this "average" is the scaled area under the partial receiver operating characteristic (ROC) curve. The ROC curve is a popular graphical representation of pairs of one minus the specificity and sensitivity as the cutoff value varies (Pepe (2003)). An appealing property of ROC-based criteria is their independence of the prevalence rate of the event of interest. Consequently, the ROC curve can be estimated consistently in a case−control study, where fixed numbers of cases and controls are selected according to the plan of the researcher, which may not reflect their true proportions in the general population.

However, there seems to be a disconnect between the construction of a scoring system and its evaluation standard. When the regression model at the first stage is correctly specified, the resulting feature combination is automatically optimal for many of the criteria used at the second stage (Jin and Lu (2009)). However, in the most general case, where the regression model does not contain the true model, this two-stage approach may be too generic, thus yielding suboptimal solutions for the specific criterion of interest. Ideally, if an evaluation criterion is important and is used to evaluate the quality of the scoring system, we should construct the scoring system based on the same criterion. Following this line of reasoning, Pepe, Cai and Longton (2006) proposed creating an ensemble of features that directly maximize the area under the ROC curve (AUC), which is a popular

method used to evaluate a scoring system (Zhou (2002); Pepe (2003); Englemann, Hayden and Tasche (2003); Blochlinger and Leippold (2006); Ravi and Pramodh (2008); Van Gool et al. (2011)). To overcome the numerical difficulties associated with the discontinuity of the objective functions, Ma and Huang (2007a) and Zhao et al. (2012) proposed maximizing a smoothed AUC by replacing the indicators with sigmoid or other surrogates. Furthermore, Komori and Eguchi (2010); Ma and Huang (2007b), and Ye et al. (2007) combined the ROC-based ensemble and various regularization approaches to construct a scoring system from many features. Ricamato and Tortorella (2011) and Wang and Chang (2011) proposed similar approaches for optimizing the area under partial ROC curves. The aforementioned methods are all model-free in the sense that the target function to be maximized approximates the area under an ROC or a partial ROC curve without needing a parametric model assumption. Assuming that the case and control features follow distinct multivariate Gaussian distributions, Su and Liu (1993); Hsu and Hsueh (2013), and Hsu, Chang and Hsueh (2014) proposed maximizing a model-based estimate of the area under an ROC or a partial ROC curve. However, the target function associated with the area under a partial ROC curve behaves badly with multiple local maximizers, including those proposed by Hsu and Hsueh (2013); Hsu, Chang and Hsueh (2014); Ricamato and Tortorella (2011) and Wang and Chang (2011). As a result, there is no reliable numerical algorithm that identifies the global optimum. Furthermore, the asymptotic properties of the estimated combinations are difficult to study. To appreciate the difficulty, consider the simple problem of finding a scoring system $S = \beta' Z$ from a feature Z that maximizes the sensitivity $P(S \geq d_0|Y = 1)$, where the specificity $P(S \leq d_0|Y = 0) = \pi_0$ for a given π_0. Both the objective and the constraint of the empirical version of this optimization problem involve a discontinuous piecewise constant function, and the optimization is often numerically intractable when the dimension of Z is greater than three. We are not aware of an existing method that solve this problem effectively. If we aim to maximize the area under the partial ROC curve, the associated optimization problem becomes even more complicated.

In this paper, we propose a new approach to creating an ensemble of features that optimizes the sensitivity for given specificity level(s) or the area under a partial ROC curve by appropriately modifying the objective and constraint functions. The target function is well behaved and the related optimization problem can be solved efficiently and reliably. In Section 2, we present the proposed approach and the associated statistical inference procedures. We perform

an extensive simulation study to investigate the operating characteristics of the proposed method and present the results in Section 3. In Section 4, we apply the proposed method to two real-data examples, in which we predict the quality of wine and "good" credit. Lastly, we conclude the paper in Section 5.

2. Method

2.1. Sensitivity-based ensemble

2.1.1. Point estimator and numerical algorithm

Suppose the observed data consist of $\{(Z_i, Y_i), i = 1, \ldots, N\}$, where Z_i and Y_i represent the feature vector and the binary response for the ith subject, respectively. Let $S_i = \beta' Z_i$ be the risk score. Our objective is to identify the optimal score by solving the following constrained optimization problem:

$$\text{maximize the sensitivity:} \quad N_1^{-1} \sum_{i=1}^{N} Y_i I(\beta' Z_i \geq d), \tag{2.1}$$

$$\text{subject to:} \quad N_0^{-1} \sum_{i=1}^{N} (1 - Y_i) I(\beta' Z_i \leq d) \geq \pi_0, \tag{2.2}$$

where $I(\cdot)$ is the indicator function and $N_0 = \sum_{i=1}^{N} (1 - Y_i)$ and $N_1 = \sum_{i=1}^{N} Y_i$ are the numbers of controls and cases, respectively. Note that the inequality constraint (2.2) can be replaced by the approximate equality constraint:

$$N_0^{-1} \sum_{i=1}^{N} (1 - Y_i) I(\beta' Z_i \leq d) \approx \pi_0.$$

This constraint holds true when the optimal scores $\{S_i, i = 1, \ldots, N\}$ have no ties. Otherwise, we can further lower the cutoff value d without reducing the corresponding sensitivity.

As discussed in the introduction, it is difficult to find the optimal weight $\hat{\beta}_{opt}$ and cutoff value $\hat{d}_{opt}$ by directly solving the optimization problem given in (2.1) and (2.2). Therefore, we present an alternative characterization for the optimal solution. To this end, we let $(\hat{\beta}_w', \hat{d}_w)'$ represent the maximizer of

$$\sum_{i=1}^{N} Y_i I(\beta' Z_i \geq d) + w \sum_{i=1}^{N} (1 - Y_i) I(\beta' Z_i \leq d), \tag{2.3}$$

and $\hat{w}$ be the solution to the equation

$$N_0^{-1}\sum_{i=1}^{N}(1-Y_i)I(\hat{\beta}_w' Z_i \le \hat{d}_w) - \pi_0 \approx 0. \tag{2.4}$$

In Appendix A, we show that $\hat{\beta}_{\hat{w}}$ is an approximate solution to the original optimization problem; that is, $\hat{\beta}_{\hat{w}} \approx \hat{\beta}_{opt}$. This equivalence is not surprising because maximizing (2.3) is equivalent to minimizing the weighted misclassification error

$$\sum_{i=1}^{N} Y_i I(\beta' Z_i - d < 0) + w\sum_{i=1}^{N}(1-Y_i)I(-\beta' Z_i + d < 0), \tag{2.5}$$

for given w, which balances the trade-off between false-positive and false-negative errors. Therefore, we can find the optimal weights by minimizing the weighted loss function (2.5) and solving the corresponding univariate equation (2.4).

However, it is still not feasible to minimize the weighted misclassification error directly, owing to the discontinuity of the indicator function. In the following, we propose solving a simpler optimization problem by replacing the indicator function $I(x<0)$ with a convex surrogate $g(x)$ (Hastie and Zhu (2006)). Here, we choose $g(x) = \log\{1+\exp(-x)\}$, and the surrogate loss function becomes

$$l_w(\beta, d) = \sum_{Y_i=1}\log\{1+\exp(d-\beta' Z_i)\} + w\sum_{Y_i=0}\log\{1+\exp(-d+\beta' Z_i)\}, \tag{2.6}$$

which is the negative of the log-likelihood function from the weighted logistic regression model

$$P(Y_i = 1|Z_i) = \frac{\exp(-d+\beta' Z_i)}{1+\exp(-d+\beta' Z_i)},$$

with all controls weighted by w. Therefore, $\hat{\beta}_w$ is simply the maximum likelihood estimator of the weighted logistic regression. Furthermore, under this framework, $I(\beta' Z_i < d)$ is approximated by $P(Y_i = 0|Z_i) = \{1+\exp(-d+\beta' Z_i)\}^{-1}$, which also allows us to construct the smoothed counterpart of equation (2.4), as follows:

$$N_0^{-1}\sum_{i=1}^{N}(1-Y_i)\{1+\exp(-d+\beta' Z_i)\}^{-1} = \pi_0. \tag{2.7}$$

The new constraint is smooth and often substantially improves the finite-sample performance of the estimated combination. In summary, we propose using $\hat{\beta}_S = \hat{\beta}(\hat{w})$ as optimal weights when combining the features, where

$$\begin{cases} \{\hat{\beta}(w), \hat{d}(w)\} = \operatorname{argmin}_{(\beta,d)} l_w(\beta, d), \\ N_0^{-1}\sum_{i=1}^{N}(1-Y_i)\{1+\exp(-\hat{d}(\hat{w}) + \hat{\beta}'(\hat{w})Z_i)\}^{-1} = \pi_0. \end{cases} \tag{2.8}$$

In general, $\hat{\beta}_S$ differs from $\hat{\beta}_{opt}$. However, the resulting risk score $S_i = \hat{\beta}_S' Z_i$ may still exhibit satisfactory sensitivity because it indirectly maximizes a surrogate function of the sensitivity. To compute $\hat{\beta}_S$, we use the following algorithm:

1. Fixing the current $\hat{\beta}$, minimize the loss function $l_w(\hat{\beta}, d)$ with respect to d. Let the solution be $\hat{d}(w)$. Furthermore, let $\hat{w}$ be the root of the equation
$$N_0^{-1} \sum_{i=1}^{N} (1 - Y_i)\{1 + \exp(-\hat{d}(w) + \hat{\beta}' Z_i)\}^{-1} = \pi_0 \quad \text{and} \quad \hat{d} = \hat{d}(\hat{w}).$$
2. Fixing the current $(\hat{d}, \hat{w})$, find $\hat{\beta}$ by minimizing $l_{\hat{w}}(\beta, \hat{d})$ with respect to β.
3. Repeat steps 1 and 2 until convergence. Let the limit of $\hat{\beta}$ be $\hat{\beta}_S$. Then, the risk score can be constructed as $S = \hat{\beta}_S' Z$ for a future subject with covariate Z. The cutoff value corresponding to the specificity of π_0, $\hat{d}_S$, is given as the π_0-th quantile of $\{\hat{\beta}_S' Z_i \mid Y_i = 0\}$. Note that, in general, $\hat{d}_S$ differs from the limit of $\hat{d}$ in steps 1 and 2.

If the sample size $N = N_0 + N_1 \to \infty$, with $N_0/N = p_0 \in (0, 1)$, we can show that $\hat{\beta}_S$ converges to a deterministic limit β_S in probability under mild regularity conditions, where $(\beta_S, \tilde{d}_S, d_S, w_S)$ is the solution of the limiting estimating equation $s_0(\beta, \tilde{d}, d, w) = 0$ given in (1) of Appendix B.

2.1.2. Confidence interval and credible set

Note that, in general, β_S and d_S do not have a meaningful interpretation with respect to the association between the feature vector Z_i and the binary outcome of interest and, thus, may not be of direct interest. However, they may serve as anchors to quantify the variability of the estimated weights and cutoff values. Specifically, we can show that $\sqrt{N}(\hat{\beta}_S' - \beta_S', \hat{d}_S - d_S)'$ converges weakly to a Gaussian distribution with mean zero and variance−covariance matrix Σ_S, which can be used to construct the confidence intervals for β_S and d_S. A direct estimation of Σ_S involves difficult nonparametric smoothing, and may be sensitive to the choice of related tuning parameters. Alternatively, one can estimate Σ_S using the resampling method. Specifically, let $(\beta_S^{*\prime}, \tilde{d}_S^*)'$ be the minimizer of
$$l_{w_S^*}^*(\beta, d) = \sum_{Y_i=1} B_i \log\{1 + \exp(d - \beta' Z_i)\} + w_S^* \sum_{Y_i=0} B_i \log\{1 + \exp(-d + \beta' Z_i)\}$$
with respect to β and d, and choose the weight w_S^* such that
$$N_0^{-1} \sum_{i=1}^{N} B_i (1 - Y_i)\{1 + \exp(-\tilde{d}_S^* + Z_i' \beta_S^*)\}^{-1} = \pi_0,$$

where the weights $\{B_i, i = 1, 2, \ldots, N\}$ are independent and identically distributed (i.i.d.) random variables from the unit exponential distribution. Lastly, let the perturbed cutoff value d_S^* be the root of the estimating equation

$$N_0^{-1} \sum_{i=1}^{N} B_i (1 - Y_i) I(Z_i' \beta_S^* \le d) \approx \pi_0.$$

Note that, in general, d_S^*, the counterpart of $\hat{d}_S$, differs from $\tilde{d}_S^*$, the counterpart of the limit of $\hat{d}$.

Conditional on the observed data, we can obtain a large number of realizations of $(\beta_S^{*\prime}, d_S^*)'$ by repeatedly generating different sets of random weights B_i and solving the constrained optimization problem. The empirical variance–covariance matrix of $\sqrt{N}(\beta_S^{*\prime} - \hat{\beta}_S', d_S^* - \hat{d}_S)'$ can then be used to estimate Σ_S. The justification is given in Appendix B. This resampling method is a special version of bootstrap method and has been used successfully in various settings (Foster, Tian and Wei (2001); Jin, Ying and Wei (2001)). Compared with the conventional bootstrap method, the independence between the random weights B_is simplifies the theoretical justification.

Furthermore, we may be interested in making an inference on the true sensitivity and specificity corresponding to the obtained $\hat{\beta}_S$ and $\hat{d}_S$. Note that both the sensitivity $P(\hat{\beta}_S' Z_i \ge \hat{d}_S | \hat{\beta}_S, \hat{d}_S, Y_i = 1)$ and the specificity $P(\hat{\beta}_S' Z_i \le \hat{d}_S | \hat{\beta}_S, \hat{d}_S, Y_i = 0)$ depend on the estimators $\hat{\beta}_S$ and $\hat{d}_S$ and, thus, are random variables. It is obvious that the sensitivity and specificity converge to $\eta_0 = P(\beta_S' Z_i \ge d_S | Y_i = 1)$ and π_0, respectively, as the sample size goes to infinity, owing to the consistency of $\hat{\beta}_S$ and $\hat{d}_S$. However, it is still important to quantify the uncertainty in a finite sample to and construct, for example, the credible regions. To this end, we specify the large sample approximation

$$\begin{pmatrix} P(\hat{\beta}_S' Z_i \ge \hat{d}_S | \hat{\beta}_S, \hat{d}_S, Y_i = 1) - \hat{\eta}_0 \\ P(\hat{\beta}_S' Z_i \le \hat{d}_S | \hat{\beta}_S, \hat{d}_S, Y_i = 0) - \pi_0 \end{pmatrix} \approx - \begin{pmatrix} N_1^{-1} \sum_{i=1}^{N_1} Y_i \{ I(\beta_S' Z_i \ge d_S) - \eta_0 \} \\ N_0^{-1} \sum_{i=1}^{N_0} (1 - Y_i) \{ I(\beta_S' Z_i \le d_S) - \pi_0 \} \end{pmatrix},$$

where

$$\hat{\eta}_0 = N_1^{-1} \sum_{i=1}^{N} Y_i I(\hat{\beta}_S' Z_i \ge \hat{d}_S) \quad \text{and} \quad \eta_0 = P(\beta_S' Z_i \ge d_S | Y_i = 1).$$

It is obvious that $\sum_{i=1}^{N_1} Y_i \{ I(\beta_S' Z_i \ge d_S)$ and $\sum_{i=1}^{N_0} (1 - Y_i) \{ I(\beta_S' Z_i \le d_S)$ follow independent binomial distributions, $\mathrm{B}(N_1, \eta_0)$ and $\mathrm{B}(N_0, \pi_0)$, respectively. Thus, we have

$$P \left\{ P(\hat{\beta}_S' Z_i \ge \hat{d}_S | \hat{\beta}_S, \hat{d}_S, Y_i = 1) \in I_{\alpha, \eta} \right\} \approx 1 - \alpha$$

and

$$P\left\{P(\hat{\beta}_S' Z_i \le \hat{d}_S | \hat{\beta}_S, \hat{d}_S, Y_i = 0) \in I_{\alpha,\pi}\right\} \approx 1 - \alpha$$

for large N, where

$$I_{\alpha,\eta} = \left[2\hat{\eta}_0 - \frac{c_{1-\alpha/2,B(N_1,\hat{\eta}_0)}}{N_1}, 2\hat{\eta}_0 - \frac{c_{1-\alpha/2,B(N_1,\hat{\eta}_0)}}{N_1}\right],$$
$$I_{\alpha,\pi} = \left[2\pi_0 - \frac{c_{1-\alpha/2,B(N_0,\pi_0)}}{N_0}, 2\pi_0 - \frac{c_{1-\alpha/2,B(N_1,\pi_0)}}{N_0}\right],$$

and $c_{\alpha,B(N,p)}$ is the α-quantile of the binomial distribution $\mathrm{B}(n,p)$. Therefore, $I_{\alpha,\eta}$ and $I_{\alpha,\pi}$ are the $(1-\alpha)$ credible intervals for the sensitivity and specificity, respectively. Furthermore, $\hat{\Omega}_\alpha = I_{\alpha/2,\eta} \times I_{\alpha/2,\pi}$ can serve as the joint credible region of the potential sensitivity and specificity if we apply the constructed score and estimated cutoff value to future patients from the same population. The theoretical justification is given in Appendix B.

Remark 1. The choice of $g(\cdot)$ is not unique. For example, we can let $g(x) = \exp(-x)$ or the hinge loss $(1-x)_+ = \max(0, 1-x)$. Although these alternative choices cannot fit into a coherent statistical model, such as, a logistic regression, constraints (2.4) or (2.7) can still be coupled with the new objective function to yield efficient algorithms that combine multiple features.

Remark 2. A sufficient condition to ensure the convergence of the proposed algorithm is that

$$\sum_{i=1}^{N}(1-Y_i)\left[1+\exp\{\hat{\beta}(w)'Z_i - \hat{d}(w)\}\right]^{-1}$$

is a monotone function of the weight w. This condition is almost always satisfied in practice, because we can show that a similar function,

$$\sum_{i=1}^{N}(1-Y_i)\log\left(\left[1+\exp\{\hat{\beta}(w)'Z_i - \hat{d}(w)\}\right]^{-1}\right),$$

is monotone in w. The justification is given in Appendix B.

Remark 3. The cutoff value associated with the estimated risk score $\hat{\beta}_S' Z_i$ is given by the π_0-th quantile of the observed scores of all controls. This may be similar to $\hat{\tilde{d}}_S$, the root of the estimating equation

$$N_0^{-1}\sum_{i=1}^{N}(1-Y_i)\left\{1+\exp(-d+\hat{\beta}_S' Z_i)\right\}^{-1} = \pi_0.$$

However, we prefer $\hat{d}_S$ to $\hat{\tilde{d}}_S$ because the former ensures that the observed sensi-

tivity level is π_0 without needing an approximation or a model assumption.

2.2 Partial ROC curve-based combination

2.2.1. Point estimator and numerical algorithm

In practice, rather than a single specificity level π_0, we may be interested in, for example, all specificity levels within a given interval $[\pi_L, \pi_U]$. In such a case, the predictiveness of the risk score S is measured as the area under the partial ROC curve over the interval $[1-\pi_U, 1-\pi_L]$. For observed data, the area under the partial empirical ROC curve is

$$\int_{\pi_L}^{\pi_U} N_1^{-1} \sum_{i=1}^{N} Y_i I\{S_i \geq \hat{d}(\pi)\} d\pi, \tag{2.9}$$

where $\hat{d}(\pi)$ satisfies the equality

$$N_0^{-1} \sum_{i=1}^{N} (1-Y_i) I\{S_i \leq \hat{d}(\pi)\} \approx \pi.$$

Similarly to the discussion in Section (2.1), the weight of the best linear combination $S_i = \beta' Z_i$ that maximizes (2.9) can be approximated by $\hat{\beta}_R = \hat{\beta}_{\hat{w}(\cdot)}$, where

$$\begin{cases} \{\hat{\beta}_{\hat{w}(\cdot)}, \hat{d}_{\hat{w}(\cdot)}(\cdot)\} = \operatorname{argmin}_{\{\beta, d(\cdot)\}} \int_{\pi_L}^{\pi_U} l_{\hat{w}(\pi)}\{\beta, d(\pi)\} d\pi, \\ N_0^{-1} \sum_{i=1}^{N} (1-Y_i) \left[1 + \exp\left\{-\hat{d}_{w(\cdot)}(\pi) + \hat{\beta}'_{w(\cdot)} Z_i\right\}\right]^{-1} = \pi, \end{cases} \tag{2.10}$$

Then as in (2.8), we can use the following algorithm to solve this constrained optimization problem:

1. For given $\hat{\beta}$, minimize the loss function $l_w(\hat{\beta}, d)$ with respect to d, and denote the minimizer by $\hat{d}(w)$. Furthermore, for any $\pi \in [\pi_L, \pi_U]$, let $\hat{w}(\pi)$ be the root of the estimating equation

$$N_0^{-1} \sum_{i=1}^{N} (1-Y_i) \left[1 + \exp\{-\hat{d}(w) + \hat{\beta}' Z_i\}\right]^{-1} = \pi.$$

2. For given $\hat{w}(\pi)$ and $\hat{d}(w)$, find $\hat{\beta}$ by minimizing

$$\int_{\pi_L}^{\pi_U} l_{\hat{w}(\pi)}[\beta, \hat{d}\{\hat{w}(\pi)\}] d\pi.$$

3. Repeat steps 1 and 2 until convergence. Let the limit of $\hat{\beta}$ be $\hat{\beta}_R$. Finally, the risk score is constructed as $S = \hat{\beta}'_R Z$ for a future subject, with covariate Z. The cutoff value associated with the specificity π is $\hat{d}_R(\pi)$, the π-the

quantile of $\{\hat{\beta}_R' Z_i \mid Y_i = 0\}$.

2.2.2. Confidence interval and credible set

In Appendix C, we show that $\hat{\beta}_R$ converges to a deterministic limit β_R in probability under mild regularity conditions. Here $\{\beta_R, \tilde{d}_R(\pi), d_R(\pi), w_R(\pi), \pi \in [\pi_L, \pi_U]\}$ is the solution to the functional estimating equation $m_0\{\beta, \tilde{d}(\cdot), d(\cdot), w(\cdot)\}(\pi) = 0, \pi \in [\pi_L, \pi_U]$ given in (2) in Appendix C. Furthermore, we can show that $\sqrt{N}\{\hat{\beta}_R' - \beta_R', \hat{d}_R(\pi_1) - d_R(\pi_1), \ldots, \hat{d}_R(\pi_K) - d_R(\pi_K)\}', \pi_k \in [\pi_L, \pi_U]$, for $k = 1, 2, \ldots, K$, converges weakly to a normal distribution with zero mean and a variance−covariance matrix Σ_R, where K is a given integer. A direct estimation of Σ_R is difficult. Thus we can estimate Σ_R using a resampling method similar to that introduced in Section 2.1. Specifically, let $\{\beta_R^*, \tilde{d}_R^*(\cdot)\}$ be the minimizer of

$$\int_{\pi_L}^{\pi_U} l^*_{\pi_R^*(\pi)}\{\beta, d(\pi)\} d\pi$$

with respect to $\{\beta, d(\cdot)\}$, where the weight function $w_R^*(\cdot)$ is chosen such that

$$N_0^{-1} \sum_{i=1}^{N} B_i(1 - Y_i)[1 + \exp\{-\tilde{d}_R^*(\pi) + Z_i' \beta_R^*\}]^{-1} = \pi, \pi \in [\pi_L, \pi_U],$$

where the weights $\{B_i, i = 1, 2, \ldots, N\}$ are i.i.d. random variables from the unit exponential distribution. Lastly, we select $d_R^*(\cdot)$ satisfying

$$N_0^{-1} \sum_{i=1}^{N} B_i(1 - Y_i) I\{Z_i' \beta_R^* \le d_R^*(\pi)\} \approx \pi, \pi \in [\pi_L, \pi_U].$$

Conditional on the observed data, we can obtain a large number of realizations of $\{\beta_R^*, d_R^*(\cdot)\}$ by repeatedly generating different sets of random weights B_i and solving the constrained optimization problem. Then, we can use the empirical variance−covariance matrix of $\sqrt{N}\{\beta_R^{*\prime} - \hat{\beta}_R', d_R^*(\pi_1) - \hat{d}_R(\pi_1), \ldots, d_R^*(\pi_K) - \hat{d}_R(\pi_K)\}'$ to estimate Σ_R. The justification is similar to that for $\hat{\beta}_S$, given in the Appendix B.

We may also want to know the true area under the partial ROC curve based on the estimated risk score $\hat{\beta}_R' Z$ in the future population, that is,

$$\int_{\pi_L}^{\pi_U} P\left\{\hat{\beta}_R' Z_i \ge \hat{d}_R(\pi) \mid Y_i = 1, \hat{\beta}_R, \hat{d}_R(\cdot)\right\} d\pi.$$

Because the true area under the partial ROC curve depends on $\hat{\beta}_R$ and $\hat{d}_R(\cdot)$, it is a random variable. To construct a credible interval for this variable with a

desired probability, we employ the approximation

$$\begin{aligned}&\sqrt{N_1}\left[\hat{\tau}_R-\int_{\pi_L}^{\pi_U}P\left\{\hat{\beta}_R'Z_i\geq\hat{d}_R(\pi)\mid Y_i=1,\hat{\beta}_R,\hat{d}_R(\cdot)\right\}d\pi\right]\\ \approx&\sqrt{N_1}\left[\int_{\pi_L}^{\pi_U}\frac{1}{N_1}\sum_{Y_i=1}I\{\beta_R'Z_i\geq d_R(\pi)\}d\pi-\tau_R\right],\end{aligned}$$

where

$$\tau_R=\int_{\pi_L}^{\pi_U}P\left\{\beta_R'Z_i\geq d_R(\pi)\mid Y_i=1\right\}d\pi$$

and

$$\hat{\tau}_R=\int_{\pi_L}^{\pi_U}N_1^{-1}\sum_{Y_i=1}I\{\hat{\beta}_R'Z_i\geq\hat{d}_R(\pi)\}d\pi.$$

Furthermore, we generate many realizations of

$$W=\int_{\pi_L}^{\pi_U}N_1^{-1}\left[\sum_{i=1}^{N_1}I\{U_i\leq\hat{\eta}_R(\pi)\}\right]d\pi$$

by repeatedly simulating $\{U_j,j=1,\ldots,N_1\}$ from the uniform distribution $U[0,1]$ to approximate the distribution of

$$\int_{\pi_L}^{\pi_U}N_1^{-1}\sum_{Y_i=1}I\{\beta_R'Z_i\geq d_R(\pi)\}d\pi,$$

where $\hat{\eta}_R(\pi)=N_1^{-1}\sum_{Y_i=1}I\{\hat{\beta}_R'Z_i\geq\hat{d}_R(\pi)\}$ is a consistent estimator of $\eta_R(\pi)=P\{\beta_R'Z_i\geq d_R(\pi)\}$. This approximation is based on the fact that $I\{\beta_R'Z_i\geq d_R(\pi)\}$ and $I\{U_i\leq\eta_R(\pi)\}$, for $\pi\in[\pi_L,\pi_U]$, have the same distribution. Therefore,

$$\lim_{N\to\infty}P\left(\int_{\pi_L}^{\pi_U}P\left\{\hat{\beta}_R'Z_i\geq\hat{d}_R(\pi)\mid Y_i=1,\hat{\beta}_R,\hat{d}_R(\cdot)\right\}d\pi\in I_{\alpha,\eta}\right)=1-\alpha$$

and

$$I_{\alpha,\eta}=\left[2\hat{\tau}_R-\frac{c_{1-\alpha/2,W}}{N_1},2\hat{\tau}_R-\frac{c_{\alpha/2,W}}{N_1}\right]$$

serves as an asymptotic valid credible set for $\int_{\pi_L}^{\pi_U}P\{\hat{\beta}_R'Z_i\geq\hat{d}_R(\pi)\mid Y_i=1,\hat{\beta}_R,\hat{d}_R(\cdot)\}d\pi$, where $c_{\alpha,W}$ is the estimated α-th quantile of W, based on the aforementioned Monte Carlo simulation.

Remark 4. Because the area under the partial ROC curve can be approximated by

$$(\pi_U-\pi_L)\times\frac{1}{N_1K}\sum_{k=1}^{K}\sum_{i=1}^{N}Y_iI\{S_i\geq\hat{d}(\tilde{\pi}_k)\}d\pi,$$

for $\pi_L = \tilde{\pi}_1 < \tilde{\pi}_2 < \cdots < \tilde{\pi}_K = \pi_U$, in practice, we can minimize the objective function

$$\sum_{k=1}^{K}\sum_{i=1}^{N} l_{\tilde{w}_k}(\beta, d_k)$$

under the constraints

$$N_0^{-1}\sum_{i=1}^{N}(1-Y_i)\left\{1+\exp(-d_k+\beta' Z_i)\right\}^{-1} = \tilde{\pi}_k, k = 1, \ldots, K$$

for selected $\tilde{\pi}_k$.

Remark 5. When the dimension of Z_i is high relative to the sample size N, the regularization method can be easily adapted for the proposed framework. Consider the sensitivity-oriented combination as an example. Here, we can employ the popular lasso method to select informative features by modifying step 2 of the algorithm described in Section 2.1.1 (Tibshirani (1996)):

> Fixing the current $\hat{d}$ and $\hat{w}$, find $\hat{\beta}$ by minimizing
>
> $$l_{\hat{w}}(\beta, \hat{d}) + \lambda|\beta|_1,$$
>
> where $|\beta|_1$ is the L_1-norm of the vector β and λ is the penalty parameter.

The lasso-penalty parameter λ controls the sparsity of the final solution $\hat{\beta}_S(\lambda)$ or $\hat{\beta}_R(\lambda)$ and can be selected using the data-dependent cross-validation method. Then, the objective function in step 2 is convex, and the regularized minimization can often be performed by modifying the existing algorithm. Specifically, for $g(x) = \log\{1 + \exp(-x)\}$, the optimization is equivalent to fitting a lasso-regularized logistic regression model with a known intercept, and the associated numerical algorithm (e.g., the coordinate descending algorithm) is well developed.

3. Numerical Study

3.1. Simulation design

We performed extensive simulation to investigate the operational characteristics of the proposed method in a finite sample. To give an overview of the results in this section, we examined the following:

(a) the ability of the proposed method in finding feature combinations with high sensitivity at a fixed specificity level;

(b) the empirical performance of the resampling method-based inference proce-

dures when characterizing the uncertainty of the optimal weights $\hat{\beta}_S$;

(c) the empirical coverage level of the proposed credible set for the true sensitivity;

(d) the ability to identify informative features in a moderately high-dimensional case.

To this end, we simulated a training set consisting of an equal number of cases and controls (i.e., $N_1 = N_0 = 200$). The p-dimensional feature vector Z_i for the cases and controls are generated from different distributions. Specifically, we let $Z_i \sim N(\mu_0, \Sigma_0)$ and $N(\mu_1, \Sigma_1)$ for the controls and cases, respectively, where various choices of μ_0, μ_1, Σ_0, and Σ_1 are considered. Specifically, we considered the following eight cases:

1. $p = 2$, $\mu_0 = (0, 0)'$, $\mu_1 = (1, 0)'$,
$$\Sigma_0 = \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix};$$
2. $p = 2$, $\mu_0 = \mu_1 = (0, 0)'$,
$$\Sigma_0 = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 1 \end{pmatrix};$$
3. $p = 3$, $\mu_0 = (0, 0, 0)'$, $\mu_1 = (0, 0, 1)'$,
$$\Sigma_0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} 1 & -0.80 & 0.64 \\ -0.80 & 1 & -0.80 \\ 0.64 & -0.80 & 1 \end{pmatrix};$$
4. As in case 3, except $\mu_1 = (1, 0, 1)'$;
5. $p = 4$, $\mu_0 = (0, 0, 0, 0)'$, $\mu_1 = (1, 0, 0, 0)'$,
$$\Sigma_0 = \begin{pmatrix} 1 & 0.80 & 0.64 & 0.51 \\ 0.80 & 1 & 0.80 & 0.64 \\ 0.64 & 0.80 & 1 & 0.80 \\ 0.51 & 0.64 & 0.80 & 1 \end{pmatrix}, \quad \Sigma_1 = \begin{pmatrix} 1 & -0.80 & 0.64 & -0.51 \\ -0.80 & 1 & -0.80 & 0.64 \\ 0.64 & -0.80 & 1 & -0.80 \\ -0.51 & 0.64 & -0.80 & 1 \end{pmatrix};$$
6. As in case 4, except $\mu_1 = (1, 0, 1, 0)'$;
7. As in case 4, except $\mu_1 = (1, 1, 1, 0)'$;
8. As in case 4, except $\mu_1 = (1, 1, 1, 1)'$.

For the first case, the variance−covariance of the biomarkers from the cases is the same as that from the controls, and only the first biomarker is informative. In this case, a single biomarker Z_{i1} is the optimal choice for maximizing the ROC curve at any specificity level, and the simple logistic regression is expected to perform well. In this ideal setting for a logistic regression, we plan to investigate the potential loss in efficiency of the new method. In the second case, although there is no difference in the mean level, the variance−covariance of the biomarkers differs between the cases and controls. Under this setting, the optimal discriminant function $Z_{i1}Z_{i2}$ is nonlinear, and the optimal linear combination depends on the targeted specificity level. Here, we examine whether the proposed method improves the performance of the simple logistic regression, which we expect will find it difficult to identify a high-quality combination at some specificity levels. Cases 3 and 4 investigate more complicated and realistic settings, where both the mean and the covariance structure of the biomarkers depend on the outcome. The optimal discriminant function consists of both linear and quadratic components and varies with the specificity level. Cases 5−8 represent similar settings, but consider more biomarkers with higher correlations between them.

To further summarize the various cases, we plot the ROC curves based on the optimal discriminant function and the linear combination from fitting the simple logistic regression; see Figures 1 and 2 of Appendix D. It is clear that the simple logistic regression is far from optimal for cases 2−8 (especially for cases 2 and 8). These models are designed to examine whether our simple method can take advantage of the suboptimality of the standard logistic regression, which is not sensitive to differences between the covariance structures.

3.2. Simulation results

In the first set of studies, we have examined the true sensitivity and specificity of combinations of multiple features estimated using different methods. The targeted specificity level π_0 was set at 95%. For comparison purposes, we implemented three methods: our proposed method, the logistic regression, and a grid search to directly maximize the sensitivity. When implementing the proposed method, the maximum likelihood estimator from the regular logistic regression was used as the initial value for β. The grid search was only used for $p = 2$ and 3, as the method is very time consuming for $p \geq 4$. In the grid search, we reparametrized the weights β as $(\cos\theta_1, \sin\theta_1)'$, for $\theta_1 \in [0, \pi]$ and $(\cos\theta_1, \sin\theta_1\cos\theta_2, \sin\theta_1\sin\theta_2)'$, for $(\theta_1, \theta_2)' \in [0, \pi] \times [0, 2\pi]$, for $p = 2$ and 3,

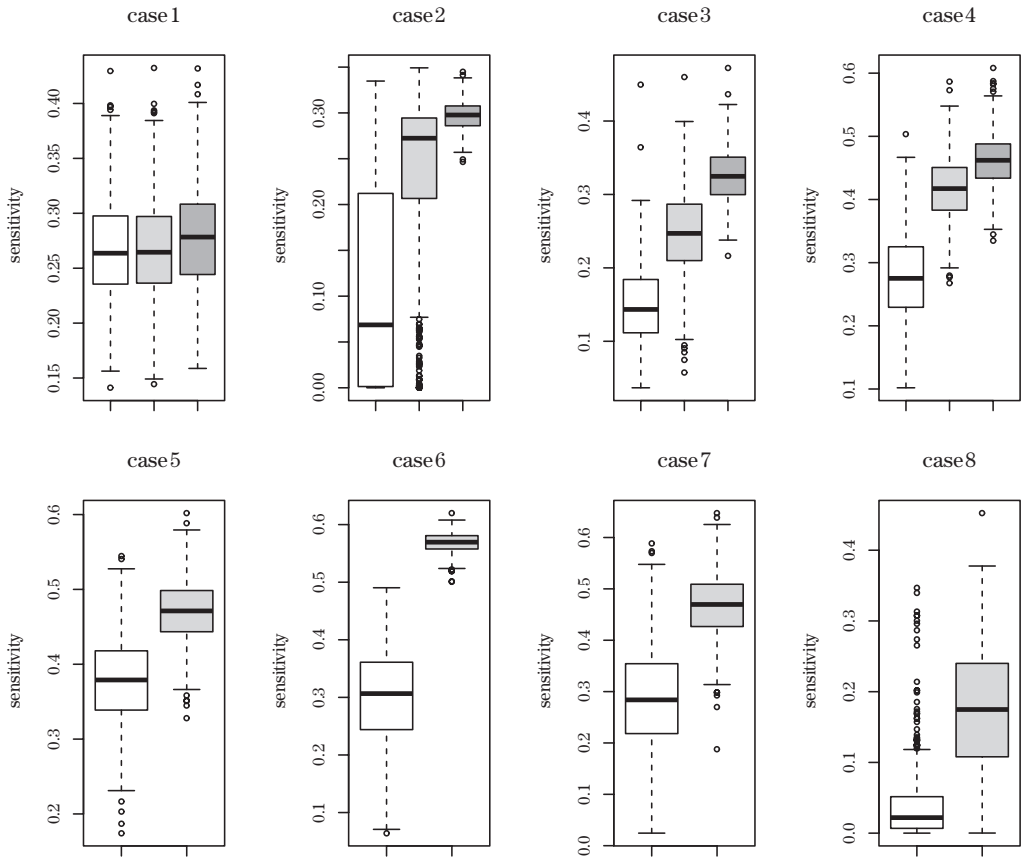

Figure 1. Boxplots for the empirical distributions of the realized sensitivities of the risk score constructed using three methods: white, logistic regression; light gray, proposed method; dark gray, grid search.

respectively, as proposed by Hsu, Chang and Hsueh (2014). The true sensitivity and specificity of the resulting combinations were estimated using an independently generated validation set of $50,000$ cases and controls. Repeating the simulation 500 times, we then compared the "true" sensitivities and specificities in the validation sets.

Figure 1 plots the empirical distributions of the realized sensitivities in the validation set. In cases $2-4$, the grid search that optimizes the empirical sensitivity in the training set always yields the best result, as expected. The second best results are achieved by the proposed method, which is sometimes substantially better than the logistic regression. In case 1, all three methods perform similarly, suggesting that the proposed method is comparable with the logistic regression,

Table 1. The empirical performance of the resampling method: bias, empirical bias; ESE, empirical standard error; ASE, empirical average of the estimated standard error; COV, the empirical coverage probability.

case	β_S	bias	ESE	ASE	COV
3	−0.052	−0.010	0.127	0.131	95.8%
	0.076	0.014	0.125	0.133	96.0%
	1.019	0.049	0.173	0.162	92.4%
4	0.929	0.040	0.158	0.164	93.6%
	0.238	0.027	0.164	0.164	95.4%
	0.928	0.037	0.170	0.164	92.6%
6	0.832	0.007	0.177	0.178	94.8%
	−0.808	0.018	0.186	0.189	95.2%
	1.017	0.020	0.189	0.192	95.6%
	−0.332	−0.013	0.176	0.181	96.4%
7	0.210	0.032	0.241	0.229	92.6%
	1.321	0.063	0.313	0.289	91.2%
	0.888	0.028	0.215	0.217	94.4%
	−1.351	−0.041	0.222	0.212	94.0%
case	d_S	bias	ESE	ASE	COV
3	1.675	0.032	0.186	0.184	93.6%
4	2.198	0.032	0.223	0.216	93.4%
6	1.383	0.006	0.128	0.142	97.0%
7	2.590	0.043	0.295	0.270	92.2%

which is the optimal choice. When the grid search becomes infeasible, the performance of the proposed method continues to be superior to that of the logistic regression, particularly in case 8. Although all three methods aim to control the specificity level at 95%, the true average specificity from the grid search tends to be slightly lower than 95%. The true specificities from the other two methods are above 94% in all cases.

In the second set of simulations, we studied the performance of the proposed resampling method. For each simulated training set, we obtained $\hat{\beta}_S$, variance estimates based on the resampling method, and the corresponding Wald−type 95% confidence intervals. To approximate the true β_S, we simulated 100,000 independent cases and controls and treated the corresponding estimator of β_S as the true value. Then, we examined the accuracy of the variance estimator from the resampling method and the empirical coverage level of the constructed confidence interval. To save space, we only reported the results from cases 3,4,6, and 7 in Table 1. The other cases are similar. In all of the reported cases, the empirical average of the estimated standard errors is fairly close to the empirical standard error of $\hat{\beta}_S$. Furthermore, the empirical coverage level of the 95%

Table 2. The empirical probabilities of selecting informative and noise features using lasso-logistic regression and the proposed sensitivity-based ensemble method. The empirical averages of the realized sensitivity (sen.) are also reported.

method		Emp. Prob. of Being Selected					sen.
	ρ	Z_1	Z_2	Z_3	Z_4	Noise markers	
Logistic reg.		100%	100%	99%	72%	13%	0.62
Sensitivity-based	0.0	100%	100%	87%	39%	6%	0.60
Logistic reg.		100%	100%	100%	74%	11%	0.77
Sensitivity-based	0.4	100%	100%	91%	46%	5%	0.76
Logistic reg.		100%	100%	91%	65%	9%	0.89
Sensitivity-based	0.8	92%	99%	75%	38%	4%	0.88

confidence interval is almost the same as the nominal level. Similarly, we also examine the variance estimates of $\hat{d}_S$ and the performance of the corresponding 95% confidence interval for d_S. The results are also satisfactory (Table 1).

In the third set of simulations, we investigated whether the proposed credible sets have an appropriate coverage level for the true sensitivity and the specificity of the estimated combination. To this end, we estimated the credible intervals and the true values of both the sensitivity and the specificity from the simulated training and validation sets. The empirical coverage level based on 500 simulations was recorded for each simulation setup. The results were summarized in Figure 2. The empirical coverage levels of the credible sets for the sensitivity and specificity are very close to the nominal level in all eight cases.

In the fourth set of simulations, we specifically examined the performance of the proposed method in terms of selecting informative features. To this end, we considered a moderately high-dimensional case with the following covariance matrix of the auto-regressive structure: $\Sigma_0 = \Sigma_1 = \{\rho^{|i-j|}\}_{p\times p}$, for $p = 100$. The mean vectors of the controls and cases are $\mu_0 = (0, \ldots, 0)'$ and $\mu_1 = \Sigma_0(1, 1, 1/2, 1/4, 0, \ldots, 0)'$, respectively. In this setting, the optimal combination of features is a linear combination of the first four features, with weights of 1, 1, 1/2, and 1/4, respectively. The optimality of this combination holds for any combination of sensitivity and specificity. We applied the lasso-regularized logistic regression and the proposed method in order to estimate the optimal linear combination with $\pi_0 = 90\%$. In Table 2, we summarizes the empirical probabilities of being selected for both informative and noise features as well as the resulting sensitivities. For all settings examined, the informative features are slightly more likely to be identified by the lasso-regularized logistic regression than they are by the proposed method. However, the logistic regression also selects more noise features. For example, when $\rho = 0$, the logistic regression

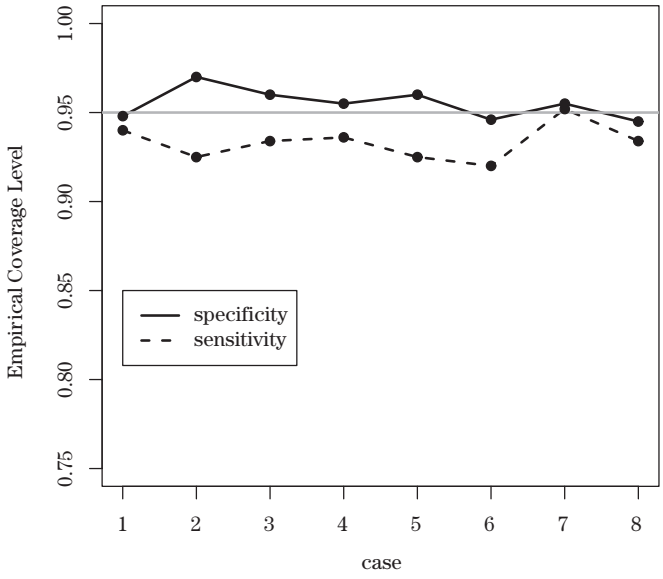

Figure 2. Empirical coverage levels for the constructed credible sets of sensitivity and specificity: solid line, the coverage probability for specificity; dashed line, the coverage probability for sensitivity.

mistakenly chooses 13 noise features, on average, compared to six by the proposed method. As a result, the sensitivities of the combinations from the logistic regression and the proposed method are very close: 62% vs. 60%. The slight superiority of the logistic regression is expected, because the logistic regression is the true model under this specific simulation design.

We also investigated the parallel properties of the partial ROC curve-based combinations, and obtained similar results; see Appendix E.

4. Example

In the first example, we tested our proposed method on the "white wine" data set studied by Cortez et al. (2009). The data set contains measurements for 4,898 white wine samples and is available in the University of California, Irvine (UCI), Machine Learning Repository. Each of the wine samples has been evaluated by at least three wine experts for quality, summarized on a scale from 0 to 10, with 0 and 10 representing the poorest and highest quality, respectively. In addition, 11 physicochemical features, including fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol were measured for all samples. Cortez et al. (2009) compared the ability of different data−mining methods to predict the ordinal

quality measurement based on the 11 physicochemical features, concluding that the support vector machine gave the most promising results. Here, we conducted a simpler analysis to identify wine samples with quality above six. These wines are considered superior and account for only 21.6% of all samples in the data set. To this end, we coded $Y = 1$ if the quality is ≤ 6, and 0 otherwise. Suppose we attempt to screen out a high proportion of good wine samples, that is, with a high specificity level π_0, by combining the 11 features. Setting $\pi_0 = 95\%$, we applied the proposed method in order to maximize the sensitivity level. The resulting weights of the features are reported in Table 3. The sensitivity of the score is 38.9%, which is 28% higher than the sensitivity of 30.3% achieved by the logistic regression. In Figure 3, we plotted the ROC curves of the scores constructed from the proposed method and the general logistic regression. Although the two curves have similar areas under the curve ($\approx 79\%$), the ROC curve based on the proposed method is clearly superior to that based on the logistic regression for the high-specificity region, but sacrifices its performance for low specificities. In this example, the approximated specificity

$$N_0^{-1} \sum_{i=1}^{N} (1 - Y_i)[1 + \exp\{-\hat{d}(w) + \hat{\beta}'(w) Z_i\}]^{-1}$$

is a smooth monotone increasing function of w. Solving the corresponding estimating equation suggests that each sample of good wine should be weighted by 58 to correspond to a specificity level of 95%. Furthermore, one may construct the credible intervals for the true specificity and sensitivity as [93.7%, 96.4%] and [37.3%, 40.5%], respectively. We also constructed the confidence interval for β_S based on the proposed resampling method. Here, we find that the contributions to the combination from the following features are statistically significantly different from zero: fixed acidity, volatile acidity duration, citric acid, residual sugar, density, pH, sulphates, and alcohol.

Next, we applied the proposed method to construct a scoring system that optimizes the area under the partial ROC curve corresponding to the specificities within [0.85, 0.95]. The resulting score is fairly similar to that using $\hat{\beta}_S$ with $\pi_0 = 95\%$ (Table 2). The achieved area under the partial ROC curve is 0.047 with a 95% credible interval of [0.046, 0.049]. That is again substantially higher than that of the logistic regression model, which is 0.043 [0.041, 0.044]. If we optimize the area under the partial ROC curve corresponding to the specificities within [0.50, 0.95], the resulting weights are, in general, between those of logistic regression and $\hat{\beta}_S$, with $\pi_0 = 95\%$ as their compromise (Table 3). The

Table 3. Estimated weights of standardized physicochemical features (with unit standard deviation) for discriminating "good" and "poor" white wine samples. All weights are normalized such that "fixed acidity" has an unit weight.

features	logistic reg.	$\hat{\beta}_S$	$\hat{\beta}_R$	$\hat{\beta}_R$
specificity		0.95	$(0.85, 0.95)$	$(0.50, 0.95)$
fixed acidity	1.000	1.000	1.000	1.000
volatile acidity	−0.819	−1.384	−1.293	−1.077
citric acid	−0.192	−0.270	−0.272	−0.244
residual sugar	3.214	3.287	3.211	3.188
chlorides	−0.593	−0.166	−0.214	−0.397
free sulfur dioxide	0.316	−0.090	−0.023	0.142
total sulfur dioxide	−0.025	0.244	0.192	0.086
density	−4.231	−3.823	−3.780	−3.954
pH	1.083	0.919	0.889	0.943
sulphates	0.531	0.438	0.440	0.472
alcohol	0.376	0.635	0.602	0.484

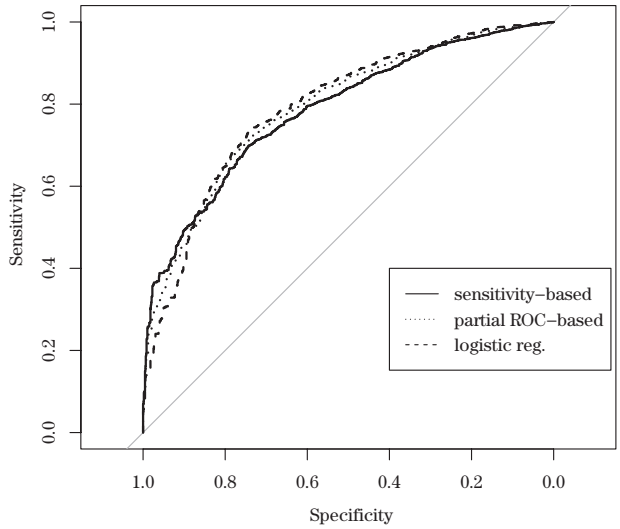

Figure 3. The ROC curves of the constructed risk scores for classifying white wine quality: solid, the proposed method, optimizing the sensitivity at the specificity level of 95%; dotted, the proposed method, optimizing the area under the partial ROC curve corresponding to specificities within [0.50, 0.95]; dashed, the logistic regression.

corresponding ROC curve, with an area under the partial ROC of 0.306 [0.301, 0.311], lies between those based on the logistic regression and $\hat{\beta}_S$ (Figure 3). The difference between the estimated weights highlights the need to apply a relevant criterion when constructing the scoring system.

In our second example, we applied the proposed method to a German credit

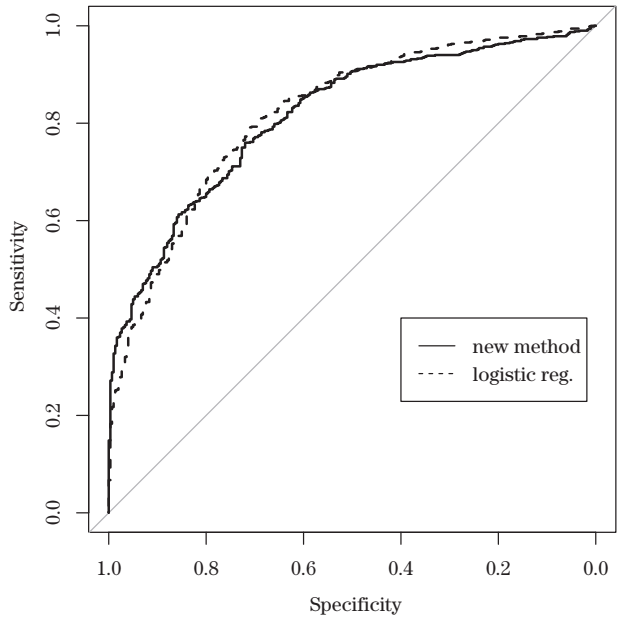

Figure 4. The ROC curves of the constructed risk scores for discriminating between credit defaults: solid, the proposed method optimizing the sensitivity at the specificity level of 95%; dashed, the logistic regression.

data set, which is also available in the UCI Machine Learning Repository. The data set consists of 20 features: checking account status, duration of the credit, credit history, purpose of the credit, credit amount, saving account status, current employment time, installment rate, marriage status, co-applicant, guarantor, time at the present residence, property, age, other installment plans, housing status, number of existing credits, type of job, number of dependents, telephone, and foreign worker. Each observation is rated as "good" or "bad", which is the binary outcome variable of interest. The data set includes 300 records of "bad" credit and 700 records of "good" credit. Numerous classification methods, including the simple logistic regression, support vector machine, and Bayesian network, have been applied to the credit data, with different degrees of success (Kuhn and Johnson (2013)). Our objective is to construct a risk score that differentiates between "good" credit and "bad" credit. Because it is worse to classify a "bad" credit as "good" than to classify a "good" credit as "bad" (from the perspective of a bank), we label "good" as one and "bad" as zero. We also set a high specificity level of 95% in the feature ensemble, thus, guaranteeing that we identify 95% of the "bad" credits. Then, we applied the proposed method to maximize the sensitivity level. To this end, credit duration, age, and time at the present resi-

dence were log-transformed; marriage status was categorized as "male/divorced," "male/single," "male/married," and "females" and the purpose of the credit was grouped into "new car purchasing," "used car purchasing," "appliance/repairs/ education/training", and "business and others". For simplicity, ordinal features were treated as numerical, for example, the five categories, "unemployed," "<1 year," "1 to 4 years," "4 to 7 years," and ">7 years" were coded as 1, 2, 3, 4, and 5, respectively, for the feature "current employment time."

The sensitivity of the resulting score is 44.7% [40.9%, 48.6%], which is non-trivially higher than the sensitivity of 38.6% [34.9%, 42.3%] achieved by the logistic regression. In Figure 4, we plot the ROC curves of the scores constructed using the proposed method and the standard logistic regression. The ROC curve based on the new method is clearly above that from the logistic regression over the region of high specificity levels. The contributions to the combination from the following features are statistically significantly different from zero: checking account status, credit duration, credit history, credit amount, time of present employment, other installment plans, foreign work, purpose of the credit, and marriage status. We also applied the proposed method to construct a scoring system that optimizes the area under the partial ROC curve corresponding to specificities between 0.85 and 0.95. The achieved area under the partial ROC curve is 0.053, with a 95% confidence interval of [0.049, 0.056]. This is higher than that of the logistic regression model, which has a value of 0.048 [0.045, 0.052].

5. Discussion

Based on our study, the optimal feature ensemble with respect to the sensitivity at a given specificity level or the area under the partial ROC curve can be differ significantly from that optimizing the area under the entire ROC curve. Therefore, it is important to select an appropriate objective function matching the most relevant criterion when combining multiple features. We proposed a novel approach for combining multiple features for binary classifications, aimed at optimizing the sensitivity or the area under the partial ROC curve. Compared with existing approaches, we do not attempt to maximize a ill-behaved target function directly. Instead, we try to solve an appropriately modified constrained optimization problem, in which the objective function is smooth and convex and has an appealing connection with the weighted logistic regression. It is possible to introduce regularization to deal with high−dimensional features under the

same framework. In a survival analysis, the c-index is a natural generalization of the area under the ROC curve. Generalizing the concept of the area under the partial ROC curve and the development of an optimization procedure similar to that presented here is left to future research.

Supplementary Material

The supplementary material is available at `http://www3.stat.sinica.edu.tw/statistica/`. Appndexi A, B, and C provide a theoretical justification of the proposed method. Appendix D provides further illustrations of the simulation setup, and Appendix E reports the simulation results for partial ROC curve-based method.

Acknowledgements

We thank the editor, AE, and two reviewers for their careful reading and constructive comments. Dr. Lu Tian's research is partially supported by R01 HL089778-08 from the National Institutes of Health, USA.

References

Blochlinger, A. and Leippold, M. (2006). Economic benefit of powerful credit scoring. *Journal of Banking and Finance* **30**, 851–873.

Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78**, 1–3.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T. and Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems* **47**, 547–553.

Englemann, B., Hayden, E. and Tasche, D. (2003). Testing rating accuracy. *Credit Risk* **16**, 82–86.

Foster, A., Tian, L. and Wei, L. (2001). Estimation for box-cox transformation model without assuming parametric error distribution. *Journal of American Statistical Association* **96**, 1097–1101.

Hastie, T. and Zhu, J. (2006). Discussion of "Support vector machine with applications" by Javier Moguerza and Alberto Munoz. *Statistical Science* **21**, 352–357.

Hsu, M., Chang, Y. and Hsueh, H. (2014). Biomarker selection for medical diagnosis using the partial area under the roc curve. *BMC Research Notes* **7**, 25.

Hsu, M. and Hsueh, H. (2013). The linear combination of biomarkers which maximize the partial area under the ROC curves. *Computational Statistics* **28**, 647–666.

Jin, H. and Lu, Y. (2009). The optimal linear combination of multiple predictors under the generalized linear models. *Statistics and Probability Letters* **79**, 2321–2327.

Jin, Z., Ying, Z. and Wei, L. (2001). A simple resampling method by perturbing the minimand. *Biometrika* **88**, 381–390.

Komori, O. and Eguchi, S. (2010). A boosting method for maximzing the partial area under the ROC curve. *BMC Bioinformatics* **11**, 314.

Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*. Springer, New York.

Ma, S. and Huang, J. (2007a). Combining multiple markers for classification using ROC. *Biometrics* **63**, 751–757.

Ma, S. and Huang, J. (2007b). Regularized ROC method for disease classification and biomarker selection with microarray data. *Bioinformatics* **21**, 4356–4362.

Pepe, M. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press.

Pepe, M., Cai, T. and Longton, G. (2006). Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics* **62**, 221–229.

Ravi, V. and Pramodh, C. (2008). Threshold accepting trained principla component neural network and feature subset selection: Application to bankruptcy predcition in banks. *Applied Soft Computing* **8**, 1539–1548.

Ricamato, M. and Tortorella, F. (2011). Partial AUC maximization in a linear combiantion of dichotomizers. *Pattern Recognization* **44**, 2669–2677.

Su, J. and Liu, J. (1993). Linear combinations of multiple diagnostic markers. *Journal of American Statistical Association* **88**, 1350–1355.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **58**, 267–288.

Van Gool, J., Verbeke, W., Sercu, P. and Baesens, B. (2011). Credit scoring for microfinance: is it worth it? *International Journal of Finance and Economics* **17**, 103–123.

Wang, Z. and Chang, Y. (2011). Marker selection via maximizing the partial area under the ROC curve of linear risk scores. *Biostatistics* **12**, 369–385.

Wilson, P., D'Agostino, R., Levy, D., Belanger, A., Silbershatz, H. and Kannel, W. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation* **97**, 1837–1847.

Ye, J., Liu, H., Kirmiz, C., Lebrilla, C. and Rocke, D. (2007). On the analysis of glycomics mass spectrometry DAA via the regularized area under the ROC curve. *BMC Bioinformatics* **8**, 477.

Zhao, X., Chen, B., Xie, Y., Tian, M., Liu, H. and Liang, X. (2012). Variable selection using the optimal ROC curve: An application to a traditional Chinese medicine study on osteoprosis disease. *Statistics in Medicine* **31**, 628–635.

Zhou, X. (2002). *Statistical Methods in Diagnostic Medicine*. John Wiley & Sons.

Department of Finance, Guanghua School of Management, Peking University, Beijing, China.

E-mail: zheng86@gsm.pku.edu.cn

Department of Biomedical Data Science, Stanford University, Palo Alto, CA 94305, USA.

E-mail: Ying.Lu@va.gov

Department of Biomedical Data Science, Stanford University, Palo Alto, CA 94305, USA.

E-mail: lutian@stanford.edu

(Received September 2014; accepted November 2017)