

Statistica Sinica Preprint No: SS-2014-0057.R2

| | |
|--|---|
| Title | On Feature Ensemble Optimizing the Sensitivity and Partial ROC Curve |
| Manuscript ID | SS-2014-0057.R2 |
| URL | http://www.stat.sinica.edu.tw/statistica/ |
| DOI | 10.5705/ss.202014.0057 |
| Complete List of Authors | Zheng Zhang Ying Lu and Lu Tian |
| Corresponding Author | Zheng Zhang |
| E-mail | zheng86@gsm.pku.edu.cn |
| Notice: Accepted version subject to English editing. | |

On Feature Ensemble Optimizing the Sensitivity and Partial ROC Curve

Zheng Zhang, Ying Lu and Lu Tian

Abstract: We consider a setting in which we want to construct a binary classifier from a panel of features to optimize either the sensitivity at a fixed specificity level or the area under the partial ROC curve. To this end, we propose an efficient iterative numerical algorithm to solve a simple constrained optimization problem which mimics the original target. We also present the associated asymptotical 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 datasets and show that the proposal can outperform the classifiers based on the generic logistic regression without considering the specific criterion we want to optimize. We also apply the new proposal to two real data examples.

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

1 Introduction

There often is a need to predict a binary outcome based on a collection of features. For example, the credit rating model is developed to identify the credit card defaulter. The rating system essentially is a classification tool signalling possible future status of individuals of interest. The rating score for each individual is calculated with features characterizing the borrower as well as the debt and aims to reflect the risk of default. Similarly, in a completely different setting, the Framingham risk score is constructed to find people susceptible to future cardiovascular attack based on baseline risk factors including age, gender, blood pressure, cholesterol level et al [Wilson et al., 1998]. In general, there are two important tasks related to this type of applications: (1) the development of a scoring system measuring the probability of the event occurrence or the rank thereof (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 one 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, i.e., the estimated linear combination or a transformation thereof, is then evaluated with selected criteria. A scoring system with good discriminatory power groups individuals with similar risks together and assigns higher score to individuals with higher risk. The quantitative quality measures for such a scoring system include misclassification error, Brier score and, very 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 cut-off 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 cut-off value d_0 is often selected to guarantee a given specificity level, e.g., $P(S \leq d_0|Y = 0) = \pi_0$, and the corresponding sensitivity is then used to measure the predictiveness of the risk score. Especially when the event of interest is relatively rare, one often needs to set a high specificity level, say $\pi_0 = 0.95$, to reduce the false positive rate when the classification tool is used in practice. Furthermore, instead of a single unanimously agreed specificity level, there may be a range of possible specificity levels for practical consideration, e.g., $[\pi_L, \pi_U]$. In such a case, one may want to use the “average sensitivity” corresponding to 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, which is the popular graphical presentation of pairs of one minus the specificity and sensitivity when the cut-off value is varied [Pepe, 2003]. One appealing property of the ROC-based criteria is its independence of the prevalence of the event of interest. Consequently, the ROC curve can be consistently estimated in the case-control study, where fixed numbers of cases and controls are selected according to the plan of researchers and may not reflect their true proportions in the general population.

There seems to be a disconnection between the construction of the scoring system and the evaluation standard. When the regression model at the first stage is correctly specified, the resulting feature combination is automatically optimal in many 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,

such a two-stage approach may be too generic and yields suboptimal solutions for the specific criterion of our interest. Ideally, if an evaluation criterion is important and used to evaluate the quality of the scoring system, one should construct the scoring system based on the same criterion at the first place. Along this line, Pepe et al. [2006] proposed to ensemble features to directly maximize the area under the ROC curve (AUC), which is a popular choice for evaluating a scoring system [Zhou, 2002, Pepe, 2003, Englemann et al., 2003, Blochlinger and Leipold, 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 [2007b] and Zhao et al. [2012] proposed to maximize a smoothed AUC by replacing the indicators by sigmoid or other surrogates. Furthermore, Komori and Eguchi [2010], Ma and Huang [2007a] and Ye et al. [2007] combined the ROC-based ensemble and various regularization approaches to construct the scoring system from a ultra-large number of features. Ricamato and Tortorella [2011] and Wang and Chang [2011] proposed similar approaches for optimizing area under the partial ROC curves. All the aforementioned methods are model-free in the sense that the targeting function to be maximized approximates the area under ROC or partial ROC curves without the need of any parametric model assumption. Assuming that the features of case and control follow distinct multivariate Gaussian distributions, Su and Liu [1993], Hsu and Hsueh [2013] and Hsu et al. [2014] proposed to maximize the model-based estimate for the area under ROC or partial ROC curve. However, one remaining big obstacle is that the target function associated with the area under partial ROC curve including those proposed by Hsu and Hsueh [2013], Hsu et al. [2014], Ricamato and Tortorella [2011] and Wang and Chang [2011], is ill-behaved with multiple local maximizers and there is no reliable numerical algorithm to find the global optimum. Furthermore, the asymptotical 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 the feature Z such that the sensitivity $P(S \geq d_0|Y = 1)$ is maximized, while the specificity $P(S \leq d_0|Y = 0) = \pi_0$ for a given π_0 . Both the objective and constraint of the empirical version of this optimization problem involve discontinuous piece-wise constant function and the optimization is often numerically intractable when the dimension of Z is higher

than 3. We are not aware of any existing method to effectively tackle this simple question. If we aim to maximize the area under the partial ROC curve, the associated optimization problem becomes even more complicated.

In this paper, we proposed a new approach on features ensemble aiming to optimize the sensitivity for given specificity level(s) or the area under partial ROC curve by appropriately modifying the objective as well as constraint functions. The target function involved is well behaved and the related optimization problem can be solved efficiently and reliably. In section 2, we presented the proposed approach and associated statistical inference procedures. We performed extensive simulation study to investigate the operating characteristics of the proposed method and the results were reported in section 3. In section 4, we applied the proposed method to two real data examples on predicting the wine quality and “good” credit. Lastly, we concluded the paper with some discussions in section 5.

2 Method

2.1 Sensitivity-based Ensemble

Point Estimator and Numeric Algorithm

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

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

$$\text{subject to: } N_0^{-1} \sum_{i=1}^N (1 - Y_i) I(\beta'Z_i \leq d) \geq \pi_0, \quad (2)$$

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) can be approximately replaced by an equality constraint:

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

This is true when the optimal scores $\{S_i, i = 1, \dots, N\}$ have no ties, since otherwise one always can further lower the cut-off value d without reducing the corresponding sensitivity.

As discussed in the introduction, it is very difficult to find the optimal weight $\hat{\beta}_{opt}$ and cut-off value \hat{d}_{opt} by directly solving the optimization problem given in (1) and (2). Therefore, we will present an alternative characterization for the optimal solution. To this end, we let $(\hat{\beta}'_w, \hat{d}_w)'$ represent the maximizer of

$$\sum_{i=1}^N Y_i I(\beta' Z_i \geq d) + w \sum_{i=1}^N (1 - Y_i) I(\beta' Z_i \leq d) \quad (3)$$

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

$$N_0^{-1} \sum_{i=1}^N (1 - Y_i) I(\hat{\beta}'_w Z_i \leq \hat{d}_w) - \pi_0 \approx 0. \quad (4)$$

In the Appendix A, we shown that $\hat{\beta}_{\hat{w}}$ is approximately the solution of the original optimization problem, i.e., $\hat{\beta}_{\hat{w}} \approx \hat{\beta}_{opt}$. This equivalence is not a surprise considering the fact that maximizing (3) is equivalent to minimizing the weighted misclassification error

$$\sum_{i=1}^N Y_i I(\beta' Z_i - d < 0) + w \sum_{i=1}^N (1 - Y_i) I(-\beta' Z_i + d < 0) \quad (5)$$

for given w , which balances the trade-off between false positive and false negative errors. Therefore, we may find the optimal weights by minimizing the weighted loss function (5) and solving the corresponding univariate equation (4).

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

$$l_w(\beta, d) = \sum_{Y_i=1} \log\{1 + \exp(d - \beta' Z_i)\} + w \sum_{Y_i=0} \log\{1 + \exp(-d + \beta' Z_i)\}, \quad (6)$$

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 the 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 (4) as

$$N_0^{-1} \sum_{i=1}^N (1 - Y_i) \{1 + \exp(-d + \beta'Z_i)\}^{-1} = \pi_0. \quad (7)$$

The new constraint is smooth and often can substantially improve the finite sample performance of the estimated combination. In summary, we proposed to use $\hat{\beta}_S = \hat{\beta}(\hat{w})$ as optimal weights in combining the features, where

$$\begin{cases} \{\hat{\beta}(w), \hat{d}(w)\} = \operatorname{argmin}_{(\beta, d)} l_w(\beta, d) \\ N_0^{-1} \sum_{i=1}^N (1 - Y_i) \{1 + \exp(-\hat{d}(\hat{w}) + \hat{\beta}'(\hat{w})Z_i)\}^{-1} = \pi_0 \end{cases}. \quad (8)$$

Although $\hat{\beta}_S$ is in general different from $\hat{\beta}_{opt}$, one may still expect that the resulting risk score $S_i = \hat{\beta}'_S Z_i$ has a satisfactory sensitivity since it indirectly maximizes a surrogate function of the sensitivity. To compute $\hat{\beta}_S$, we present 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 current (\hat{d}, \hat{w}) , find $\hat{\beta}$ by minimizing $l_{\hat{w}}(\beta, \hat{d})$ with respect to β .
3. Repeat steps 1-2 until convergence. Let the limit of $\hat{\beta}$ be $\hat{\beta}_S$ and the risk score can be constructed as $S = \hat{\beta}'_S Z$ for a future subject with covariate Z . The cut-off value corresponding to the specificity of π_0 , \hat{d}_S , is given as the π_0 -th quantile of $\{\hat{\beta}'_S Z_i | Y_i = 0\}$. Note that in general \hat{d}_S is different from the limit of \hat{d} in steps 1 and 2.

If the sample size $N = N_0 + N_1 \rightarrow \infty$ with $N_0/N = p_0 \in (0, 1)$, one 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.

Confidence Interval and Credible Set

Note that β_S and d_S in general do not have any meaningful interpretation with respect to the association between the feature vector Z_i and binary outcome of interest and thus may not be of our direct interest. However, they may serve as anchors to quantify the variability of the estimated weights and cutoff values. Specifically, one 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 . The direct estimation of Σ_S involves the difficult nonparametric smoothing and may be sensitive to the choice of related tuning parameters. Alternatively, one may estimate Σ_S by resampling method. Specifically, let $(\beta_S^{*'}, \tilde{d}_S^*)'$ 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 the weight w_S^* is chosen 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 $\{B_i, i = 1, 2, \dots, N\}$ are independent identically distributed random variables from the unit exponential distribution. Lastly, let the perturbed cut-off value \tilde{d}_S^* be the root of the estimating equation:

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

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

Conditional on the observed data, one may obtain a large number of realizations of $(\beta_S^{*'}, \tilde{d}_S^*)'$ by repeatedly generating different sets of random weights

B_i s and solving the constrained optimization problem. The empirical variance-covariance matrix of $\sqrt{N}(\beta_S^* - \hat{\beta}'_S, d_S^* - \hat{d}_S)'$ can then be used to estimate Σ_S . The justification is given in the Appendix B. This resampling method is a special version of bootstrap method and has been successfully used in various settings [Foster et al., 2001, Jin et al., 2001]. Compared with the conventional bootstrap, the independence between the random weights B_i s simplifies the theoretical justification.

Furthermore, one may be interested in making inference on the true sensitivity and specificity corresponding to the obtained $\hat{\beta}_S$ and \hat{d}_S . Note that both the sensitivity $P(\hat{\beta}'_S Z_i \geq \hat{d}_S | \hat{\beta}_S, \hat{d}_S, Y_i = 1)$ and specificity $P(\hat{\beta}'_S Z_i \leq \hat{d}_S | \hat{\beta}_S, \hat{d}_S, Y_i = 0)$ 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 $\eta_0 = P(\beta'_S Z_i \geq d_S | Y_i = 1)$ and π_0 , respectively, as the sample size goes to infinity due to the consistency of $\hat{\beta}_S$ and \hat{d}_S . However, it is still important to quantify the related uncertainty in finite sample and construct for example their credible regions. To this end, we note the large sample approximation

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

where

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

It is obvious that $\sum_{i=1}^{N_1} Y_i \{I(\beta'_S Z_i \geq d_S)\}$ and $\sum_{i=1}^{N_0} (1 - Y_i) \{I(\beta'_S Z_i \leq d_S)\}$ follow independent binomial distributions $B(N_1, \eta_0)$ and $B(N_0, \pi_0)$, respectively, and thus we have

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

and

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

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

Remark 1 The choice of $g(\cdot)$ is not unique. For example, one may let $g(x) = \exp(-x)$ or the hinge loss $(1 - x)_+ = \max(0, 1 - x)$. Although those alternative choices can not fit into a coherent statistical model such as logistic regression, the constraint (4) or (7) can still be coupled with the new objective function to yield efficient algorithms for combining multiple features.

Remark 2 One 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 since one can show that a very 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 the Appendix B.

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

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

We still prefer \hat{d}_S to \hat{d}_S since the former ensures that the observed sensitivity level is π_0 without any approximation or model assumption.

2.2 Partial ROC curve-based Combination

Point Estimator and Numerical Algorithm

In practice, there may not be an unanimously desired specificity level π_0 . For example, one may have potential interest in all specificity levels within a given interval $[\pi_L, \pi_U]$. In such a case, the predictiveness of the risk score S can be measured by the area under the partial ROC curve over the interval $[1 - \pi_U, 1 - \pi_L]$. With observed data, the area under the partial empirical ROC curve is

$$\int_{\pi_L}^{\pi_U} N_1^{-1} \sum_{i=1}^N Y_i I\{S_i \geq \hat{d}(\pi)\} d\pi \quad (9)$$

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

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

Similar to the discussion in section (2.1), the weight of the best linear combination $S_i = \beta' Z_i$ maximizing (9) can be approximated by $\hat{\beta}_R = \hat{\beta}_{\hat{w}(\cdot)}$, where

$$\begin{cases} \{\hat{\beta}_{\hat{w}(\cdot)}, \hat{d}_{\hat{w}(\cdot)}(\cdot)\} = \operatorname{argmin}_{\{\beta, d(\cdot)\}} \int_{\pi_L}^{\pi_U} l_{\hat{w}(\pi)}\{\beta, d(\pi)\} d\pi \\ N_0^{-1} \sum_{i=1}^N (1 - Y_i) \left[1 + \exp\left\{-\hat{d}_{\hat{w}(\cdot)}(\pi) + \hat{\beta}'_{\hat{w}(\cdot)} Z_i\right\} \right]^{-1} = \pi, \end{cases} \quad (10)$$

Similar to (8), one may employ 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 1-2 until convergence. Let the limit of $\hat{\beta}$ be $\hat{\beta}_R$. Finally, the risk score can be constructed as $S = \hat{\beta}'_R Z$ for a future subject with covariate

Z. The cut-off value associated with the specificity π is $\hat{d}_R(\pi)$, the π -the quantile of $\{\hat{\beta}'_R Z_i \mid Y_i = 0\}$.

Confidence Interval and Credible Set

In the Appendix C, we have shown that $\hat{\beta}_R$ converges to a deterministic limit β_R in probability under mild regularity conditions, where $\{\beta_R, \tilde{d}_R(\pi), d_R(\pi), w_R(\pi), \pi \in [\pi_L, \pi_U]\}$ is the solution of a functional estimating equation $m_0\{\beta, \tilde{d}(\cdot), d(\cdot), w(\cdot)\}(\pi) = 0, \pi \in [\pi_L, \pi_U]$ given in (2) of Appendix C. Furthermore, one may show that $\sqrt{N}\{\hat{\beta}'_R - \beta'_R, \hat{d}_R(\pi_1) - d_R(\pi_1), \dots, \hat{d}_R(\pi_K) - d_R(\pi_K)\}'$, $\pi_k \in [\pi_L, \pi_U], k = 1, 2, \dots, K$ converges weakly to a normal distribution with zero mean and a variance-covariance matrix Σ_R , where K is a given integer. The direct estimation of Σ_R is difficult and one may estimate Σ_R by the 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 $\{B_i, i = 1, 2, \dots, N\}$ are independent identically distributed random variables from the unit exponential distribution. Lastly we select $d_R^*(\cdot)$ satisfying that

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

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

One may interested in the true area under the partial ROC curve based on

12

the estimated risk score $\hat{\beta}'_R Z$ in the future population, i.e.

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

Since 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 covering it with a desired probability, one may note 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 may 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 $\{U_j, j = 1, \dots, N_1\}$ 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 $I\{U_i \leq \eta_R(\pi)\}, \pi \in [\pi_L, \pi_U]$ have the same distribution. Therefore

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

and

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

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

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

$$(\pi_U - \pi_L) \times \frac{1}{N_1 K} \sum_{k=1}^K \sum_{i=1}^N Y_i I\{S_i \geq \hat{d}(\tilde{\pi}_k)\} d\pi$$

for $\pi_L = \tilde{\pi}_1 < \tilde{\pi}_2 < \dots < \tilde{\pi}_K = \pi_U$, in practice, one may minimize the objective function

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

under the constraints

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

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

Remark 5 When the dimensional of Z_i is high relative to the sample size N , the regularization method can be easily adapted for the proposed framework. Take the sensitivity-oriented combination as an example, one may employ the popular lasso method for selecting informative features by modifying the step 2 of the proposed algorithm [Tibshirani, 1996]:

2 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 via the data-dependent cross-validation method. The the objective function at step 2 is convex and the regularized minimization often can be performed by modifying the existing algorithm. Specifically, with $g(x) = \log\{1 + \exp(-x)\}$, the optimization is equivalent to fitting a lasso-regularized logistic regression model with known intercept and the associated numerical algorithm such as coordinate descending algorithm is well developed.

3 Numerical Study

3.1 Simulation Design

We have performed extensive simulation study to investigate the operational characteristics of the proposed method in finite sample. To give an overview of the results in this section, we studied

- (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 procedures to characterize 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 of identifying informative features in moderately high dimensional case.

To this end, we simulated the training set consisting of equal number of cases and controls, i.e., $N_1 = N_0 = 200$, and the p -dimensional feature vector Z_i s for cases and controls were generated from different distributions. Specifically, we let $Z_i \sim N(\mu_0, \Sigma_0)$ and $N(\mu_1, \Sigma_1)$ in the controls and cases, respectively, where various choices of μ_0, μ_1, Σ_0 and Σ_1 were considered. Specifically, we have considered 8 different 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. the same as case 3, except that $\mu_1 = (1, 0, 1)'$;
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. the same as case 4, except that $\mu_1 = (1, 0, 1, 0)'$;
7. the same as case 4, except that $\mu_1 = (1, 1, 1, 0)'$;
8. the same as case 4, except that $\mu_1 = (1, 1, 1, 1)'$.

For the first case, the variance-covariance of biomarkers from cases is the same as that from controls and only the first biomarker is informative. In such a 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 logistic regression, we plan to investigate the potential loss in efficiency of the new method. For the second case, while there is no difference in mean level, the variance-covariance of biomarkers is different between 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. We are interested in examining if the new proposal can improve the performance of the simple logistic regression, which is expected to have difficulty to identify a high-quality combination at some specificity levels. Cases 3 and 4 investigated more complicated and realistic settings, where both the mean and covariance structure of the biomarkers depend on the outcome status. The optimal discriminant function consists of both linear and quadratic components and varies with the specificity levels. Cases 5-8 represent similar settings with more biomarkers under consideration and higher correlations among them.

To further summarize various cases, we have plotted the ROC curves based on the optimal discriminant function and the linear combination from fitting the simple logistic regression in Figures 1 and 2 of the 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 specific models are designed to examine if our simple proposal can take advantage of the suboptimality of the standard logistic regression, which is not sensitive to capture the difference in 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 from different methods. The targeted specificity level π_0 was set at 95%. For comparison purpose, we have implemented three methods: our proposed method, the logistic regression and grid search to directly maximize the sensitivity. In 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, since the method is very time-consuming for $p \geq 4$. In grid search, we reparametrized the weights β as $(\cos \theta_1, \sin \theta_1)'$, $\theta_1 \in [0, \pi]$ and $(\cos \theta_1, \sin \theta_1 \cos \theta_2, \sin \theta_1 \sin \theta_2)'$, $(\theta_1, \theta_2)' \in [0, \pi] \times [0, 2\pi]$ for $p = 2$ and 3, respectively, as proposed by Hsu et al. [2014]. The true sensitivity and specificity of the resulting combinations were estimated in 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 optimizing 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 sometimes is substantially better than the logistic regression. In case 1, all three methods perform similarly, suggesting that the new proposal loses very little comparable with the logistic regression, which is the optimal choice. When the grid search becomes infeasible, the performance of the new proposal continues to be superior to that of the logistic regression specially 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 other two methods are above 94% in all the cases.

In the second set of simulations, we have studied the performance of the proposed resampling method. For each simulated training set, we obtained $\hat{\beta}_S$, its

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. We then examined the accuracy of the variance estimator from the resampling method and the empirical coverage level of the constructed confidence interval. To save the space, we only reported the results from cases 3,4,6 and 7 in Table 1. Other cases are similar. In all 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% 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 have studied if the proposed credible sets have appropriate coverage level for the true sensitivity and specificity of the estimated combination. To this end, we estimated the credible intervals and the true value of both sensitivity and specificity based on simulated training and validation sets, respectively. The empirical coverage level based on 500 simulations were recorded for each simulation setup and the results were summarized in Figure 2. The empirical coverage levels of credible sets for both sensitivity and specificity are very close to the nominal level in all eight cases.

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

lasso-regularized logistic regression than the new proposal. However, the logistic regression also selects more noise features. For example, when $\rho = 0$, the logistic regression on average mistakenly chooses 13 noise features compared to 6 by the new proposal. As a result, the sensitivities for combinations from the logistic regression and new method are very close: 62% vs 60%. The slight superiority of the logistic regression is expected since the logistic regression is the true model under this specific simulation design.

We have also investigated the parallel properties of partial ROC curve-based combinations and obtained similar results, which are summarized in Appendix E.

4 Example

In the first example, we tested our proposal on the “white wine” data set studied by Cortez et al. [2009]. The data set contains measurements for 4898 white wine samples and is available in the UCI Repository. Each of the wine samples was evaluated by at least three wine experts for its quality, which was summarized on a scale from 0 to 10, with 0 and 10 representing the poorest and highest quality, respectively. 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 also measured for all samples in the laboratory. Cortez et al. [2009] compared different data mining methods aiming to predict the ordinal quality measurement from these eleven physicochemical features and concluded that the support vector machine gave the most promising results. Here we conducted a simpler analysis to identify wine samples with quality above 6, which was considered superior and accounted 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 big proportion of good wines samples, i.e, requiring a high specificity level π_0 in combining 11 features. Setting $\pi_0 = 95\%$, we then applied our proposal to maximize the sensitivity level. The resulting weights of the features were 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 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 new method is clearly superior to that from the logistic regression on the high specificity region, while sacrificing its performance on 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 indeed a smooth monotone increasing function of w and solving the corresponding estimating equation suggests that each good wine sample 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 have also constructed the confidence interval for β_S based on the proposed resampling method and found 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.

We have also applied the proposed method to construct the scoring system optimizing 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], again substantially higher than that of the logistic regression model, which is 0.043 [0.041, 0.044]. If we aim to 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 corresponding ROC curve with an area under the partial ROC of 0.306 [0.301, 0.311] was sandwiched by those based on logistic regression and $\hat{\beta}_S$ (Figure 3). The difference in estimated weights highlights the need of applying the relevant criterion in constructing the scoring system.

As second demonstrating example, we have applied the proposed method to the German credit data set, which is also available in the UCI 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 sta-

tus, 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 ranging from the simple logistic regression, support vector machine to 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 differentiating the “good” credits from the “bad” ones. Since the it is much worse to classify a “bad” credit as “good” than a “good” credit as “bad” from the perspective of the bank, we labeled the “good” as 1 and “bad” as 0 and set a high specificity level of 95% in feature ensemble, i.e., guaranteeing the identification of 95% of the “bad” credits. We then applied our proposal to maximize the sensitivity level. To this end, credit duration, age and time at the present residence were log-transformed; the marriage status was categorized into “male/divorced”, “male/single”, “male/married” and “females” and the purpose of the credits 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, e.g., the five categories, “unemployed”, “<1 year”, “1 to 4 years”, “4 to 7 years” and “>7 years” are 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 plotted the two ROC curves of scores constructed via 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 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 have also applied the proposed method to construct the scoring system optimizing 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], again higher than that of the logistic regression model, 0.048[0.045, 0.052].

5 Discussion

Based on our experience, the optimal feature ensemble with respect to the sensitivity at given specificity level or the area under the partial ROC curve could be very different from that optimizing the area under the entire ROC curve. Therefore, it is important to select the appropriate objective function matching the most relevant criterion in combining multiple features. In this paper, we have proposed a novel approach to ensemble multiple features for binary classification with the purpose of optimizing the sensitivity or the area under the partial ROC curve. Compared with existing approach, we do not attempt to directly maximize the ill-behaved target function. 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. One may conveniently introduce regularization to deal with high dimensional feature under the same framework. In 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 optimization procedure similar to those presented in the current paper warrant further research.

Supplementary Materials: As the online supplementary materials, the Appendix A, B and C with theoretical justification of the proposed method are available at <http://www3.stat.sinica.edu.tw/statistica/>. There are also further illustrations of the simulation setup in Appendix D. The simulation results for partial ROC curve-based method is reported in Appendix E.

Acknowledgement: We thank 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 National Institute of Health, USA.

Bibliography

- A. Blochlinger and M. Leippold. Economic benefit of powerful credit scoring. *Journal of banking and finance*, 30, 2006.
- G. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78, 1950.
- P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009.
- B. Englemann, E. Hayden, and D. Tasche. Testing rating accuracy. *Credit Risk*, 16:82–86, 2003.
- A. Foster, L. Tian, and LJ. Wei. Estimation for box-cox transformation model without assuming parametric error distribution. *Journal of American Statistical Association*, 96(455):1097–1101, 2001.
- T. Hastie and J. Zhu. Discussion of “Support vector machine with applications” by Javier Moguerza and Alberto Munoz. *Statistical Science*, 21, 2006.
- M. Hsu and HM. Hsueh. The linear combination of biomarkers which maximize the partial area under the roc curves. *Computational Statistics*, 28, 2013.
- M. Hsu, YI. Chang, and HM. Hsueh. Biomarker selection for medical diagnosis using the partial area under the roc curve. *BMC Research Notes*, 7(25): doi:10.1186/1756-0500-7-25, 2014.
- H. Jin and Y. Lu. The optimal linear combination of multiple predictors under the generalized linear models. *Statistics and Probability Letters*, 79(22):2321–2327, 2009.

- Z. Jin, Z. Ying, and L.J. Wei. A simple resampling method by perturbing the minimand. *Biometrika*, 88:381–390, 2001.
- O. Komori and S. Eguchi. A boosting method for maximizing the partial area under the ROC curve. *BMC Bioinformatics*, 11, 2010.
- M. Kuhn and K. Johnson. *Applied Predictive Modeling*. Springer-New York, 2013.
- S. Ma and J. Huang. Regularized ROC method for disease classification and biomarker selection with microarray data. In *Bioinformatics*, volume 21, pages 4356–4362, 2007a.
- S. Ma and J. Huang. Combining multiple markers for classification using ROC. *Biometrics*, 63, 2007b.
- M. Pepe. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, 2003.
- M. Pepe, T. Cai, and G. Longton. Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics*, 62, 2006.
- V. Ravi and C. Pramo dh. Threshold accepting trained principal component neural network and feature subset selection: Application to bankruptcy prediction in banks. *Applied Soft Computing*, 8, 2008.
- M. Ricamato and F. Tortorella. Partial AUC maximization in a linear combination of dichotomizers. *Pattern Recognition*, 44:2669–2677, 2011.
- J. Su and J. Liu. Linear combinations of multiple diagnostic markers. *Journal of American Statistical Association*, 88(424):1350–1355, 1993.
- R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society (B)*, 58:267–288, 1996.
- J. Van Gool, W. Verbeke, P. Sercu, and B. Baesens. Credit scoring for microfinance: is it worth it? *International Journal of Finance and Economics*, 17(2): 103–123, 2011.

Z. Wang and Y. Chang. Marker selection via maximizing the partial area under the ROC curve of linear risk scores. *Biostatistics*, 22, 2011.

PW. Wilson, RB. D'Agostino, D. Levy, AM. Belanger, H. Silbershatz, and WB. Kannel. Prediction of coronary heart disease using risk factor categories. *Circulation*, 18:1837–1847–STS400, 1998.

J. Ye, H. Liu, C. Kirmiz, C. Lebrilla, and D. Rocke. On the analysis of glycomics mass spectrometry data via the regularized area under the roc curve. *BMC Bioinformatics*, 12, 2007.

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

XH. Zhou. *Statistical Methods in Diagnostic Medicine*. John Wiley & Sons, 2002.

Table 1: Simulation results for evaluating the empirical performance of the resampling method based on 500 simulations: bias, empirical bias; ESE, empirical standard error; ASE, empirical average of the estimated standard error; COV, the empirical coverage probability.

| | Sensitivity-based ensemble | | | | |
|------|----------------------------|--------|-------|-------|-------|
| 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% |

Table 2: The empirical probabilities of selecting informative as well as noise features based on lasso-regularized logistic regression and the proposed sensitivity-based ensemble method. The empirical averages of the realized sensitivity (sen.) is also reported.

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

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 1: The boxplots for the empirical distributions of realized sensitivities of risk score constructed via three different methods: white, logistic regression; light gray, new proposal; dark gray, grid search.

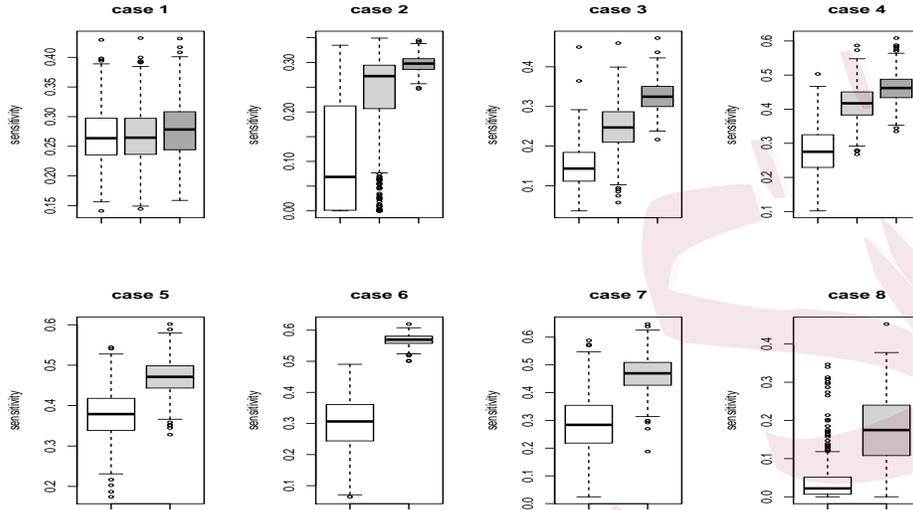


Figure 2: The empirical coverage levels for the constructed credible sets of sensitivity and specificity: solid line, the coverage probability for specificity; dashed line, the coverage probability for sensitivity

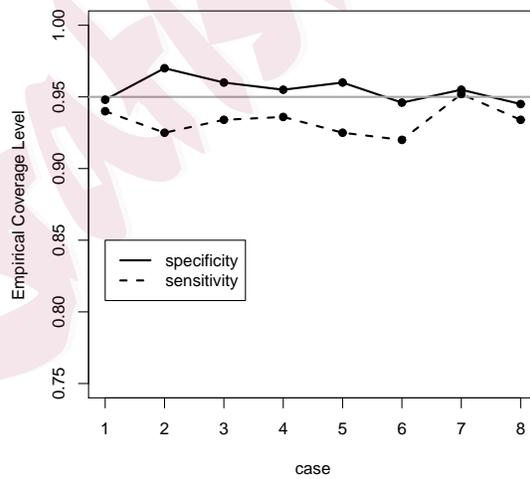


Figure 3: The ROC curves of the constructed risk scores for classifying the white wine quality: solid, the proposed method optimizing the sensitivity at the specificity level of 95%; dotted, the proposed method optimizing the area under the partial ROC curve corresponding to specificities within $[0.50,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

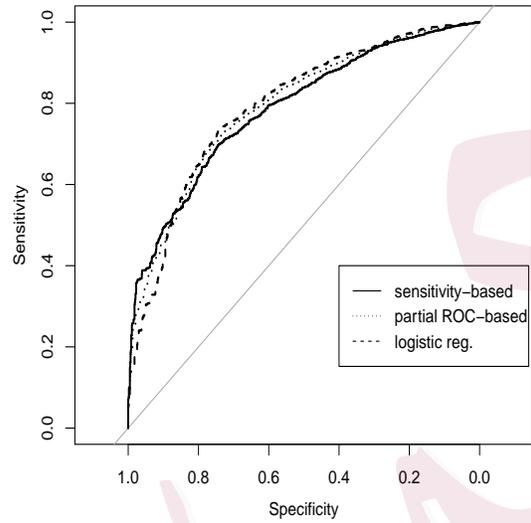
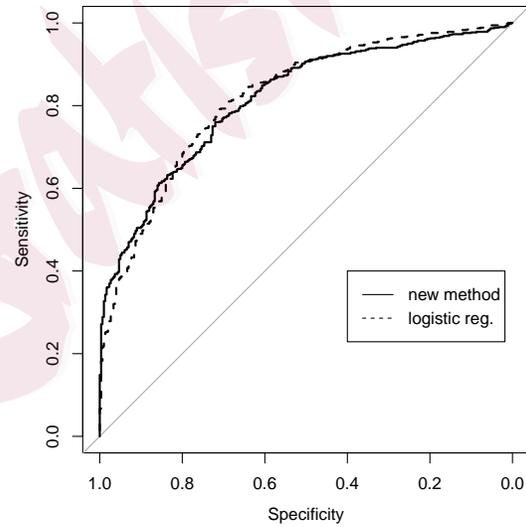


Figure 4: The ROC curves of the constructed risk scores for discriminating credit defaults: solid, the proposed method optimizing the sensitivity at the specificity level of 95%; dashed, the logistic regression.



Zheng Zhang (Corresponding Author), Department of Finance, Guanghua School of Management, Peking University, Beijing, P.R. China.

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

Ying Lu, Department of Biomedical Data Science, Stanford University and US VA Cooperative Studies Program Palo Alto Coordinating Center, Palo Alto, CA 94305, USA.

E-mail: Ying.Lu@va.gov

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

E-mail: lutian@stanford.edu