AUC based Biomarker Ensemble with an Application on Gene Scores Predicting Low Bone Mineral Density

XG. Zhao 1, W. Dai 2, Y. Li 2, and L. Tian 3*
1 Department of Bone and Joint Surgery, Xi’an School of Medicine Hospital, Xi’an, Shanxi 710077, P.R. China.
2 Department of Biostatistics, Harvard University, Boston, MA 02115
3 Department of Health Research and Policy, Stanford University, Palo Alto, CA 94301

ABSTRACT

Motivation: The area under the receiver operating characteristic (ROC) curve (AUC), long regarded as a “golden” measure for the predictiveness of a continuous score, has propelled the need to develop AUC-based predictors. However, the AUC-based ensemble methods are rather scant, largely due to the fact that the associated objective function is neither continuous nor concave. Indeed, there is no reliable numerical algorithm identifying optimal combination of a set of biomarkers to maximize the AUC, especially when the number of biomarkers is large.

Results: We have proposed a novel AUC-based statistical ensemble methods for combining multiple biomarkers to differentiate a binary response of interest. Specifically, we propose to replace the non-continuous and non-convex AUC objective function by a convex surrogate loss function, whose minimizer can be efficiently identified. With the established framework, the lasso and other regularization techniques enable feature selections. Extensive simulations have demonstrated the superiority of the new methods to the existing methods. The proposal has been applied to a gene expression data set to construct gene expression scores to differentiate elderly women with low bone mineral density (BMD) and those with normal BMD. The AUCs of the resulting scores in the independent test data set has been satisfactory.

Conclusion: Aiming for directly maximizing AUC, the proposed AUC-based ensemble methods provide an efficient means of generating a stable combination of multiple biomarkers, which is especially useful under the high dimensional settings.

Contact: lutian@stanford.edu

1 INTRODUCTION

Given that there are multiple biomarkers and a binary response of interest (e.g., case and control), it is often of substantial interest to combine the biomarkers to form a “strong” scoring system for the differentiation of cases from controls. While the choice of the predictive measure is not unique, the most appealing choice is the area under the receiver operating characteristic (ROC) curve (AUC) in the case-control study (Zhou et al., 2002; Pepe, 2003). For finite samples, AUC is simply the nonparametric two-sample Mann-Whitney U test statistics. Unlike the measures such as misclassification rate, the AUC reflects the intrinsic predictive value of a score in that it does not depend on the prevalence of the cases and thus is invariant under the case-control sampling. Therefore, it is natural to combine biomarkers by maximizing the AUC under ROC curve (Pepe et al., 2006; Ma and Huang, 2005, 2007; Ye et al., 2007; Zhou et al., 2011). However, it is notoriously difficult to maximize the AUC numerically since the objective function is neither continuous nor convex. Ad-hoc methods have been proposed to tackle the numerical problem. For example, sigmoid function has been used to approximate the indicator function used in calculating AUC (Ma and Huang, 2007). However, the smoothed objective function may still have multiple local maxima, with no guarantee of locating the global maximizer by using the commonly used numerical algorithms. In view of these challenges, we propose a class of ensemble methods aiming for maximizing AUC with multiple biomarkers. Specifically, we introduce a class of convex surrogate loss functions to approximate the non-convex AUC, greatly facilitating computation and optimization.

2 METHOD

2.1 Surrogate loss functions

Assume that \(X_1, \ldots, X_n\) are \(n\) independently identically distributed (i.i.d) copies of \(p\)-dimensional random vector \(X\), representing, for example, \(p\) biomarkers for cases, and \(Y_1, \ldots, Y_m\) are \(m\) i.i.d copies of \(p\)-dimensional random vector \(Y\) for controls. Suppose that we want to construct a score as a linear combination of the \(p\) biomarkers with the aim of maximizing the AUC under ROC curves. Specifically, we want to find a vector \(\beta\) to maximize the objective function

\[
AUC(\beta) = (nm)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} I(\beta' X_i > \beta' Y_j).
\]

Ideally we would want the score for cases to be higher than that for controls, which yields \(1\) for the AUC and completely differentiates cases and controls. However several challenges are prominent. Firstly, since the objective is invariant in \(\beta \to c \beta\) for arbitrary \(c > 0\), there is no unique maximizer for the objective function. One often needs to subjectively select an anchor biomarker with its weight in the linear combination being one and maximize the remaining \(p - 1\) components in \(\beta\). The performance of the score heavily depends on the selection of the anchor biomarker. Secondly, even with a given...
anchor biomarker the objective function is still neither continuous nor concave and therefore it is very likely that conventionally used optimization iterations have been trapped around local maximum points depending on the subjectively selected initial point.

We are now in a position to propose a method addressing these two challenges. By noting that $1 - AUC(\beta)$ can be interpreted as the misclassification rate of using the binary rule $\beta'(X_i - Y_j) > 0$ to classify a binary response always taking value 1, we may borrow the popular classification approaches aimed for minimizing the misclassification rate in the data mining literature (Hastie and Zhu, 2006; Friedman et al., 2000). Specifically, instead of directly maximizing $AUC(\beta)$ or equivalently minimizing $1 - AUC(\beta)$, it is sensible to minimize a surrogate loss function. We propose the following two surrogate functions

$$M_1(\beta) = (nm)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \log[1 + \exp(-\beta'(X_i - Y_j))],$$

and

$$M_2(\beta) = (nm)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \{1 - \beta'(X_i - Y_j)\}^+,\$$

which correspond to the negative log likelihood function raised in the conditional logistic regression and hinge loss function used in support vector machine, respectively, where $x_+ = xI(x > 0)$. Since these two functions are continuous convex function, their minimizers are well defined and can be reliably located. Numerically, it amounts to replace the indicator function $I(x < 0)$ by $\log[1 + \exp(-x)]$ and $(1 - x)_+$, respectively. Previous data mining literature reveals that algorithms minimizing such surrogate loss functions often result in models with good performance in minimizing the misclassification rate. Analogously, the estimated scores tend to render satisfactory AUCs.

With moderate dimension $p$, while $M_1(\beta)$ can be minimized via the scoring algorithm in fitting a generalized linear model, we can use linear programming technique to minimize $M_2(\beta)$. Specifically, it is equivalent to

$$\min \sum_{i=1}^{n} \sum_{j=1}^{m} \xi_{ij}$$

subject to

$$\xi_{ij} = \xi_{ij} - \xi_{ij},$$

$$\xi_{ij} = 1 - \beta'(X_i - Y_j), \quad 1 \leq i \leq n, 1 \leq j \leq m,$$

$$\xi_{ij} \geq 0, \xi_{ij} \geq 0.$$

Several advantages of the proposal are obvious. First, minimizing either $M_1(\beta)$ or $M_2(\beta)$ does not require selecting an anchor biomarker a priori, which is especially appealing for high-dimensional case. Second, in the limiting case, the maximizer of the AUC under the ROC curve

$$E\{AUC(\beta)\} = \text{pr}\{g(X) > g(Y)\}$$

is $E\{f_1(\cdot)/f_0(\cdot)\}$ for any strictly monotone increasing function $m(\cdot)$, where $f_1(\cdot)$ and $f_0(\cdot)$ are the underlying density functions of $X$ and $Y$, respectively (Jin and Lu, 2008). The minimizer of

$$E\{M_1(\beta)\} = E\log[1 + e^{-\left(g(X) - g(Y)\right)}]$$

is $\log(f_1(\cdot)/f_0(\cdot))$ and thus minimizing $E\{M_1(\beta)\}$ is equivalent to maximizing the AUC under the ROC curve. The minimizer of

$$E\{M_2(\beta)\} = E\{1 - g(X) + g(Y)\}$$

is more complicated. Numerical studies point that the minimizer may also be a monotone transformation of $f_1(\cdot)/f_0(\cdot)$, as opposed to the conventional support vector machine whose solution approximates the optimal decision boundary.

2.1.1 Adaptive Generalizations As neither $M_1(\beta)$ nor $M_2(\beta)$ provides a good approximation to $1 - AUC(\beta)$ (indeed no convex functions accurately approximate the indicator function), we employ an iterative algorithm to approximately minimize $1 - AUC(\beta)$. More specifically, given that

$$m_c(x) = \frac{(c - x)_+}{c + (c - x)_+} \rightarrow I(x < 0) \text{ as } c \rightarrow 0^+,$$

we expect that the minimizer of $\sum_{i=1}^{n} \sum_{j=1}^{m} m_c[\beta'(X_i - Y_j)]$ approximates that of $1 - AUC(\beta)$ for small $c > 0$. Figure 1 illustrates the 0-1 loss function against $\log[1 + \exp(-x)]$, the hinge loss function and $m_c(\cdot)$ with different $c$. Noting that $m_c(\beta'(x)) = m_c(\beta/c'(x))$, the scoring system minimizing $m_c(\beta'(x))$ is equivalent to that minimizing $m_1(\beta'(x))$. Thus we may employ the following adaptive algorithm

1. Set the initial

$$\beta \leftarrow \arg\min_{\beta} M_2(\beta)$$

2. Update $\beta$ as

$$\beta \leftarrow \arg\min \sum_{i=1}^{n} \sum_{j=1}^{m} \{1 - \gamma'(X_i - Y_j)\}$$

where the minimization can be solved via linear programming technique.

3. Repeat step (2) until convergence or the number of iteration reaches a pre-specified number.

Our numerical results (reported in the supplementary material) show that the adaptive iteration may increase the resulting AUC especially when there are potential outliers. The robustness of the method is not a surprise, because the influence from individual observations on the objective function via $m_c(x)$ is always bounded. Furthermore, we find that one or two iterations often suffice to harvest most of the gain in maximizing $AUC(\beta)$ and thus in general there is no need to continue the iteration until convergence.

2.1.2 Extension to Survival Outcomes When the outcome is survival time subject to potential right censoring, the $c-$ index as the generalized AUC is often computed as

$$c(\beta) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} I(\beta'Z_i > \beta'Z_j)I(\hat{T_i} < \hat{T_j})\Delta)}{\sum_{i=1}^{n} \sum_{j=1}^{m} I(\hat{T_i} < \hat{T_j})\Delta},$$

where $(\hat{T_i}, \Delta_i, Z_i), i = 1, \ldots n$ are i.i.d copies of $(T, \Delta, Z)$, $T$ is the minimum of the censoring and failure times, $\Delta$ is the binary.
censoring indicator and \( Z \) is the covariate vector (Cai and Cheng, 2008). Similarly, one may find \( \beta \) by minimizing
\[
\tilde{M}_1(\beta) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \log [1 + \exp\{-\beta'(Z_i - Z_j)\} ] I(\tilde{T}_i < \tilde{T}_j) \Delta_i,
\]
and
\[
\tilde{M}_2(\beta) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \{ 1 - \beta'(Z_i - Z_j) \} \{ I(\tilde{T}_i < \tilde{T}_j) \} \Delta_i,
\]
the counterparts of \( M_1(\beta) \) and \( M_2(\beta) \), respectively. Indeed, the log-partial likelihood function also can be viewed as surrogate to the \( c \)-index in that both
\[
\log \left[ 1 + \sum_{k=1}^{K} e^{-x_k} \right] \quad \text{and} \quad \sum_{k=1}^{K} \log(1 + e^{-x_k})
\]
may serve as a surrogate to
\[
\sum_{j=1}^{K} I(x_j < 0).
\]
This may explain that the Cox model-based \( c \)-index is often high.

\[ \text{Fig. 1. } m_c(x) \text{ against the 0-1 loss function. black: 0-1 loss function } I(x < 0); \text{ red: hinging loss function; green: } m_1(x); \text{ blue: } m_{0.5}(x); \text{ cyan: } m_{0.25}(x); \text{ purple: } \log(1 + \exp(-x)) \]

### 2.2 Regularization for high dimensional covariates

When \( p \) is high, the proposed surrogate loss function can be conveniently regularized for feature selection. While many regularization methods can be used, we hereafter pursue the popular lasso regularization for illustration. (Tibshirani, 1996). Specifically, we propose to minimize
\[
M_j(\beta) + \lambda |\beta|, \quad j = 1, 2
\]
where \( \beta = (\beta_1, \cdots, \beta_p)' \) and \( |\beta| = \sum_{j=1}^{p} |\beta_j| \). For \( M_j(\beta) \), one may use the following iterative coordinate descending algorithm to minimize the objective function with a given \( \lambda \)

1. Set an initial estimator \( \beta \)
2. Update \( \beta \)
\[
\beta \leftarrow \arg \min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} W_{ij} \{ \tilde{Z}_{ij} - \beta'(X_i - X_j) \}^2,
\]
where \( W_{ij} = p_{ij}(1 - p_{ij}) \),
\[
\tilde{Z}_{ij} = \beta'(X_i - X_j) + p_{ij}(1 - p_{ij})
\]
and
\[
p_{ij} = \frac{1}{1 + \exp\{\beta'(X_i - X_j)\}},
\]
for \( i = 1, \cdots, n \) and \( j = 1, \cdots, m \). The standard coordinate descending algorithm can be used to minimize the weighted \( L_2 \) loss function in this step (Friedman et al., 2010).
3. Repeat step 2 until convergence.

The coordinate descending algorithm is not directly applicable for the non-differentiable surrogate loss function \( M_2(\beta) \). However, since the objective function \( M_2(\beta) \) is convex and piece-wise linear, the exact solution path with \( \lambda \) varying from \( \infty \) to 0 is also piece-wise linear and can be computed using the generalized LARS algorithm (Cai et al., 2009; Rosset and Zhu, 2007). When the dimension \( p \) or sample size \( n \) is high, the computation is demanding due to dense joints in the solution path. As a remedy, we propose a forward stagewise algorithm that generates an approximate solution path of \( \beta \) (Friedman and Popescu, 2004).

1. Set an initial estimator \( \beta = 0 \) and small positive number \( \epsilon > 0 \)
2. At step \( k = 1, \cdots, \), identify the coordinate \( j \) with the largest decrease
\[
\max \{ |M_2(\beta) - M_2(\beta + e_j \epsilon)|, M_2(\beta) - M_2(\beta - e_j \epsilon) \}
\]
and update
\[
\beta \leftarrow \beta + s_j e_j \epsilon
\]
where \( \epsilon > 0 \) is a small constant selected a priori, \( e_j \) is a \( p \)-dimensional vector with all the components being zero except the \( j \)th component, which is 1 and \( s_j = 2 I \{ M_2(\beta + e_j \epsilon) < M_2(\beta - e_j \epsilon) \} - 1 \).
3. Repeat step 2 until the number of nonzero components of \( \beta \) reaches a pre-specified number or the AUC under ROC curve becomes 1.
3 SIMULATION

Extensive simulations are conducted to examine the finite sample performance of the proposed method. We generate the covariates \( \{X_1, \ldots, X_n\} \) and \( \{Y_1, \ldots, Y_n\} \) from the following models:

1. (multivariate normal) \( X_i \overset{iid}{\sim} N(\mu_1, \Sigma_1) \) and \( Y_j \overset{iid}{\sim} N(\mu_2, \Sigma_2) \), where \( X_i \) and \( Y_j \) are \( p \)-dimensional random vectors.
2. (normal mixture) \( X_i \overset{iid}{\sim} 0.8N(\mu_1, \Sigma_1) + 0.2N(\mu_3, \Sigma_3) \) and \( Y_j \overset{iid}{\sim} 0.8N(\mu_2, \Sigma_2) + 0.2N(\mu_3, \Sigma_3) \), i.e., 20\% of the markers values in both cases and controls are contaminated by a common error distribution.
3. (multivariate log-normal) \( \log(X_i) \overset{iid}{\sim} N(\mu_1, \Sigma_1) \) and \( \log(Y_j) \overset{iid}{\sim} N(\mu_2, \Sigma_2) \).
4. (log-normal mixture) \( \log(X_i) \overset{iid}{\sim} 0.8N(\mu_1, \Sigma_1)+0.2N(\mu_3, \Sigma_3) \) and \( \log(Y_j) \overset{iid}{\sim} 0.8N(\mu_2, \Sigma_2) + 0.2N(\mu_3, \Sigma_3) \).

In the above settings, we let \( \mu_1 = (1, 0, 0, \ldots, 0)' \), \( \mu_2 = (0, 1, 0, \ldots, 0)' \), \( \mu_3 = (1, 1, 1, 0, \ldots, 0)' \),

\[
\Sigma_1 = \Sigma_2 = \begin{pmatrix}
1 & 1/3 & \cdots & 1/3 \\
1/3 & 1 & \cdots & 1/3 \\
\vdots & \vdots & \ddots & \vdots \\
1/3 & 1/3 & \cdots & 1
\end{pmatrix}
\]

\( \Sigma_3 = 5I_p \). \( I_p \) is the identity matrix. We considered several configurations of \( n, m \) and \( p \) to investigate the operational characteristics of the proposed method.

Firstly, we examine the scenario where the number of covariates is low relative to the sample size. To this end, we let \( p = 3 \) and \( n = m = 50 \). For each generated data set, we construct a linear combination of the covariates as a score differentiating cases from controls, where the weights of the linear combinations are estimated by minimizing (1) \( M_1(\beta) \) (2) \( M_2(\beta) \) (3) the loss function

\[
S(\beta) = -\sum_{i=1}^n \sum_{j=1}^m \frac{1}{\sigma} + e^{-\beta'(X_i-Y_j)/\sigma}
\]

proposed in Ma and Huang (2007) and (4) by fitting a regular logistic regression. We also implement the popular ada-boosting with 300 iterations using the simple stump as the base classifier.

Furthermore, the score based on the one-step adaptive hinge loss function performs similar or slightly superior to that based on hinge loss function itself.

Secondly, we have examined the performance of the proposed method for covariates with moderate dimension. In this case, we let \( p = 200 \) and \( n = m = 50 \) and the lasso regularization is used for selecting the important features in logistic regression. The forward stagewise algorithm similar to that presented in section 2.2 for minimizing \( M_2(\beta) \) is also used to minimize \( S(\beta) \).

We choose the popular lasso penalty mainly for the purpose of fair comparison, i.e., evaluating the relative performance of various methods under similar regularization schemes. The boxplots of AUCs in independent test sets over 250 replications are plotted in Figure 3. In general, the scores based on \( M_1(\beta) \) perform better than that based on the alternatives in terms of average AUC in the test sets. Furthermore, the AUCs from scores constructed via \( M_1(\beta) \) also tend to have smaller variability than their counterparts.

In the most challenging fourth setting, the empirical average AUC in the test sets is 0.66 for score minimizing \( M_1(\beta) \), 0.66 for score minimizing the hinge loss, \( M_2(\beta) \), 0.60 for scores minimizing \( S(\beta) \), 0.60 for score from the logistic regression fitting and 0.63 for score from ada-boosting using three markers. An increase from 0.60-63 to 0.66 in the AUC is often considered nontrivial in clinical practice.

Lastly, we have examined the cases for high dimensional covariate. Here, we let \( p = 20,000 \) and \( n = m = 50 \). To save computational time, the ada-boosting is only applied to top 500 features selected based on significance levels of t-test comparing cases and controls in the training set. The simulation results are presented in Figure 4. For the high dimensional covariates, the relative performance of the proposed methods is even better than that in the previous case where \( p = 200 \). For example, in the third setting, the empirical average AUC in the test sets is 0.85 for score minimizing \( M_1(\beta) \), 0.85 for score minimizing the hinge loss, 0.76 for scores minimizing \( S(\beta) \), 0.74 for score from the logistic regression fitting and 0.64 for score from ada-boosting using three markers. Similarly, in the fourth setting, the empirical average AUC in the test sets are 0.56, 0.56, 0.53, 0.53 and 0.53 for aforementioned five methods.

4 ANALYSIS OF THE BONE MINERAL DENSITY STUDY

We apply the proposed method to a data set (Reppe et al., 2010) arising from a study that recruited 301 non-related postmenopausal ethnic Norwegian women at the Lovisenberg Deacon Hospital. Among them, bone mineral density (BMD) and gene expression levels (Affymetric array) were measured for 84 women. Since low BMD is associated with higher fracture rates (Cooper, 1997), it is of interest to identify a linear combination of gene expression levels to differentiate the osteopenia or osteoporosis(low BMD) from normal among postmenopausal women. Bone biopsies show that there are 39 from 84 women having osteopenia or osteoporosis. All the normalized gene expression level are log-transformed. After screening out approximately 25\% probesets with lowest variation,
we have 40411 probesets for each patient. We randomly split the data into training and validation sets and apply the proposed method to the training set consisting of 26 cases and 30 controls. 

For the purpose of comparisons, we also construct gene scores based (1) $S(\beta)$ proposed in Ma and Huang (2007) (2) lasso-regularized logistic regression and (3) ada-boosting using stumps as base classifiers. The anchor gene and $\sigma$ in $S(\beta)$ are determined using the same method as that presented in the simulation study. 

To save computational time, ada-boosting is only applied to the top 2000 genes according to their significance level in t-test comparing average gene expression levels between cases and controls. With the estimated scores based on $M_1(\beta)$, $M_2(\beta)$, $S(\beta)$, the regularized logistic regression and ada-boosting, we examine their corresponding AUC in the validation set. The results are shown in Figure 5. It can be seen that in general scores based on $M_1(\beta)$ and $M_2(\beta)$ yield higher AUC than that based on $S(\beta)$, the commonly used logistic regression and ada-boosting with the same number of covariates in the validation set. In Figure 5, we also plot the AUC in the training set. Since the number of covariates is much higher than the sample size, the maximum AUC (AUC=1) corresponding to complete separation between case and control in the training set is reached with 20-30 covariates for all the methods. The highest AUCs for scores based on $S_1(\beta)$, regularized logistic regression and ada-boosting are 0.764, 0.687 and 0.728, respectively, while the highest AUC is 0.708 for score based on $M_1(\beta)$ and 0.764 for score based on $M_2(\beta)$. As a reference, the AUC for age is only 0.669 in this cohort. Furthermore, while the optimal scores with $M_1(\beta)$ and $M_2(\beta)$ use 9 and 13 genes, respectively, their counterparts based $S(\beta)$, regularized logistic regression and ada-boosting use 37, 12 and 45 genes, respectively. These comparisons suggest that the genes score based $M_2(\beta)$ possesses the best combination of sparsity and prediction performance: it attains the highest AUC in the validation set with only 13 genes. The gene lists selected by these methods are heavily overlapping. For example, there are 7 common genes shared by at least three out of four linear combinations constructed based $M_1(\beta)$, $M_2(\beta)$, $S(\beta)$ and logistic regression. The probe set AFFX-M27830_M.at, which is shared by all four scores, is a member of the eight core genes reported by Reppe et al. (2010). Furthermore, gene SOST (Affymetrix ID 223869 at) shared by scores based on $M_1(\beta)$, $M_2(\beta)$ and logistic regression is also a member of the eight core genes explaining the variation of BMD and sits in the “center” of the constructed inter-molecular network sharing significant associations reported in the original paper (Reppe et al., 2010). The selected genes and their corresponding weights are summarized in Table 1, where the weights are standardized such that the probe set AFFX-M27830_M.at has the unit weight for comparison purpose. One interesting and reassuring observation is that signs of all nonzero weights were consistent across methods.

We also repeat analysis based on other random training-test splitting and obtain similar results.

Fig. 5. a) Simulation setting I (normal) b) Simulation setting II (normal mixture) c) Simulation setting III (log-normal) d) Simulation setting IV (log-normal mixture)

Fig. 3. Empirical AUC in the validation sets for different methods (A: $M_1(\beta)$; B1: $M_2(\beta)$; B2: one-step adaptive $M_2(\beta)$; C: $S(\beta)$; D: lasso regularized logistic regression; E: ada-boosting using stumps with 300 iterations) with low dimensional covariates. gray box: training set; empty box: AUC with one selected covariate; gray box: AUC with three selected covariates; dark gray box, AUC with ten selected covariates.
Table 1. The estimated scores for differentiating low and normal BMD (LR: logistic regression)

<table>
<thead>
<tr>
<th>Affymatrix ID</th>
<th>$M_1(\beta)$</th>
<th>$M_2(\beta)$</th>
<th>$S(\beta)$</th>
<th>LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFFX-M27830_M_at</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.946</td>
</tr>
<tr>
<td>211769_x_at</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.915</td>
</tr>
<tr>
<td>215887_x_at</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-0.447</td>
</tr>
<tr>
<td>217761_x_at</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.073</td>
</tr>
<tr>
<td>220900_x_at</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.915</td>
</tr>
<tr>
<td>223869_x_at</td>
<td>0.587</td>
<td>0.159</td>
<td>0</td>
<td>0.112</td>
</tr>
<tr>
<td>227405_x_at</td>
<td>0.055</td>
<td>0</td>
<td>0</td>
<td>0.024</td>
</tr>
<tr>
<td>231599_x_at</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-0.175</td>
</tr>
<tr>
<td>235102_x_at</td>
<td>0.983</td>
<td>0.113</td>
<td>0</td>
<td>2.822</td>
</tr>
<tr>
<td>237739_at</td>
<td>0.779</td>
<td>0.093</td>
<td>0</td>
<td>0.896</td>
</tr>
<tr>
<td>223820_at</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1.625</td>
</tr>
<tr>
<td>223948_at</td>
<td>1.161</td>
<td>0.185</td>
<td>0</td>
<td>2.822</td>
</tr>
<tr>
<td>206742_at</td>
<td>-0.068</td>
<td>-0.06</td>
<td>-0.026</td>
<td>0</td>
</tr>
<tr>
<td>222735_at</td>
<td>-0.639</td>
<td>-0.185</td>
<td>-0.082</td>
<td>0</td>
</tr>
<tr>
<td>244035_at</td>
<td>0.692</td>
<td>0.013</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>219747_at</td>
<td>0</td>
<td>0.007</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>235439_at</td>
<td>0</td>
<td>0.033</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>238705_at</td>
<td>0</td>
<td>0.212</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>238946_at</td>
<td>0</td>
<td>0.06</td>
<td>0.02</td>
<td>0</td>
</tr>
<tr>
<td>1552477_x_at</td>
<td>0</td>
<td>-0.02</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>206273_at</td>
<td>0</td>
<td>0</td>
<td>0.008</td>
<td>0</td>
</tr>
<tr>
<td>206307_x_at</td>
<td>0</td>
<td>0</td>
<td>0.006</td>
<td>0</td>
</tr>
<tr>
<td>206326_at</td>
<td>0</td>
<td>0</td>
<td>0.038</td>
<td>0</td>
</tr>
<tr>
<td>207369_at</td>
<td>0</td>
<td>0</td>
<td>-0.036</td>
<td>0</td>
</tr>
<tr>
<td>210045_at</td>
<td>0</td>
<td>0</td>
<td>-0.002</td>
<td>0</td>
</tr>
<tr>
<td>210174_at</td>
<td>0</td>
<td>0</td>
<td>-0.036</td>
<td>0</td>
</tr>
<tr>
<td>214412_at</td>
<td>0</td>
<td>0</td>
<td>-0.004</td>
<td>0</td>
</tr>
<tr>
<td>215196_at</td>
<td>0</td>
<td>0</td>
<td>-0.008</td>
<td>0</td>
</tr>
<tr>
<td>215431_at</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0</td>
</tr>
<tr>
<td>219566_at</td>
<td>0</td>
<td>0</td>
<td>-0.01</td>
<td>0</td>
</tr>
<tr>
<td>220554_at</td>
<td>0</td>
<td>0</td>
<td>0.09</td>
<td>0</td>
</tr>
<tr>
<td>220584_at</td>
<td>0</td>
<td>0</td>
<td>-0.012</td>
<td>0</td>
</tr>
<tr>
<td>221631_at</td>
<td>0</td>
<td>0</td>
<td>-0.01</td>
<td>0</td>
</tr>
<tr>
<td>227440_at</td>
<td>0</td>
<td>0</td>
<td>-0.078</td>
<td>0</td>
</tr>
<tr>
<td>229201_at</td>
<td>0</td>
<td>0</td>
<td>-0.108</td>
<td>0</td>
</tr>
<tr>
<td>230349_at</td>
<td>0</td>
<td>0</td>
<td>-0.01</td>
<td>0</td>
</tr>
<tr>
<td>230839_at</td>
<td>0</td>
<td>0</td>
<td>-0.024</td>
<td>0</td>
</tr>
<tr>
<td>231331_at</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0</td>
</tr>
<tr>
<td>231468_at</td>
<td>0</td>
<td>0</td>
<td>-0.112</td>
<td>0</td>
</tr>
<tr>
<td>231759_at</td>
<td>0</td>
<td>0</td>
<td>-0.106</td>
<td>0</td>
</tr>
<tr>
<td>231828_at</td>
<td>0</td>
<td>0</td>
<td>-0.004</td>
<td>0</td>
</tr>
<tr>
<td>232114_at</td>
<td>0</td>
<td>0</td>
<td>0.09</td>
<td>0</td>
</tr>
<tr>
<td>234259_at</td>
<td>0</td>
<td>0</td>
<td>0.012</td>
<td>0</td>
</tr>
<tr>
<td>234421_x_at</td>
<td>0</td>
<td>0</td>
<td>-0.036</td>
<td>0</td>
</tr>
<tr>
<td>234604_at</td>
<td>0</td>
<td>0</td>
<td>0.058</td>
<td>0</td>
</tr>
<tr>
<td>241736_at</td>
<td>0</td>
<td>0</td>
<td>0.048</td>
<td>0</td>
</tr>
<tr>
<td>243673_at</td>
<td>0</td>
<td>0</td>
<td>-0.016</td>
<td>0</td>
</tr>
<tr>
<td>243889_at</td>
<td>0</td>
<td>0</td>
<td>-0.068</td>
<td>0</td>
</tr>
<tr>
<td>244338_at</td>
<td>0</td>
<td>0</td>
<td>0.022</td>
<td>0</td>
</tr>
<tr>
<td>1553027_x_at</td>
<td>0</td>
<td>0</td>
<td>-0.008</td>
<td>0</td>
</tr>
<tr>
<td>1556803_at</td>
<td>0</td>
<td>0</td>
<td>0.044</td>
<td>0</td>
</tr>
<tr>
<td>1556938_x_at</td>
<td>0</td>
<td>0</td>
<td>0.048</td>
<td>0</td>
</tr>
<tr>
<td>1560779_x_at</td>
<td>0</td>
<td>0</td>
<td>0.032</td>
<td>0</td>
</tr>
</tbody>
</table>

Fig. 4. Empirical AUC in the validation sets for different methods (A: $M_1(\beta)$; B: $M_2(\beta)$; C: $S(\beta)$; D: lasso regularized logistic regression; E: ada-boosting using stumps with 300 iterations) with high dimensional covariates. Empty box: AUC with one selected covariate; gray box: AUC with three selected covariates; dark gray box: AUC with ten selected covariates.

Fig. 5. AUC of scores for differentiating the women having low and normal BMD. Thick lines: AUC in validation set; Thin lines: AUC in training set; red: score based on $M_1(\beta)$; green: score based on $M_2(\beta)$; blue: score based on $S(\beta)$; black: score based on logistic regression; cyan: score based on ada-boosting

5 DISCUSSION

Motivated by recent advances in data mining, we have proposed a class of methods combining biomarkers to construct a scoring system, boosting the resulting AUC under the ROC curve, a prevalence-free summarization of intrinsic predictive values of a
continuous score. The method is easily adapted to high-dimensional cases, wherein one may need to identify informative features from thousands of candidate biomarkers. In high-dimensional case, we propose to apply lasso regularization to yield a parsimonious combination maximizing the AUC. On the other hand, lasso is neither the unique nor the universally optimal regularization method for analyzing high dimensional data. Due to the convexity of the proposed loss function, it is straightforward to couple \( M_j(\beta) \) with other penalty functions such as elastic net, adaptive lasso and SCAD, which may have superior performance to simple lasso in specific settings (Zou and Hastie, 2005; Zou, 2006; Zou and Li, 2008). The key proposal is to target a convex surrogate loss function instead of a discontinuous Mann-Whitney rank statistic.

While in this paper we have focused on the hinge loss function (corresponding to the 1–norm support vector machine), our results can be extended to accommodate other versions of SVM loss functions, such as

\[
(nm)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \left(1 - \beta'(X_i - Y_j)\right)^{\alpha} 
\]

for any given \( \alpha \geq 1 \). Another alternative is the exponential function used in boosting algorithm

\[
(nm)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} e^{-\beta (X_i - Y_j)}.
\]

Finally, while the AUC under the entire ROC curve is a useful global measure, the AUC under partial ROC curve has recently emerged as a useful problem-specific measure in practice (Komori and Equchi, 2010). Therefore an numerically efficient algorithm combining multiple biomarkers to maximize the AUC under partial ROC curves or sensitivity for given specificity level is worth further investigations.

6 ACKNOWLEDGEMENT

This work has been supported by R01 HL089778-04 (L. Tian) and R01 CA95747 (Y. Li) from National Institute of Health.

REFERENCES