Published in final edited form as:

Stat Med. 2016 September 20; 35(21): 3792–3809. doi:10.1002/sim.6956.

Combining Biomarkers Linearly and Nonlinearly for Classification Using the Area Under the ROC Curve

Yuyi Fong^{*,\$,#}, Shuxin Yin^{*,#}, and Ying Huang^{*,\$,*}

^{*}Fred Hutchinson Cancer Research Center, Public Health Sciences, 1100 Fairview Avenue N., M2-B500, Seattle, WA 98109, USA

^{\$}Department of Biostatistics, University of Washington, Seattle, WA, 98195, USA

Abstract

In biomedical studies, it is often of interest to classify/predict a subject's disease status based on a variety of biomarker measurements. A commonly used classification criterion is based on AUC - Area under the Receiver Operating Characteristic Curve. Many methods have been proposed to optimize approximated empirical AUC criteria, but there are two limitations to the existing methods. First, most methods are only designed to find the best linear combination of biomarkers, which may not perform well when there is strong nonlinearity in the data. Second, many existing linear combination methods use gradient-based algorithms to find the best marker combination, which often result in sub-optimal local solutions. In this paper, we address these two problems by proposing a new kernel-based AUC optimization method called Ramp AUC (RAUC). This method approximates the empirical AUC loss function with a ramp function, and finds the best combination by a difference of convex functions algorithm. We show that as a linear combination method, RAUC leads to a consistent and asymptotically normal estimator of the linear marker combination when the data is generated from a semiparametric generalized linear model, just as the Smoothed AUC method (SAUC). Through simulation studies and real data examples, we demonstrate that RAUC out-performs SAUC in finding the best linear marker combinations, and can successfully capture nonlinear pattern in the data to achieve better classification performance. We illustrate our method with a dataset from a recent HIV vaccine trial.

Keywords

AUC; biomarker combination; classification; kernel; ramp loss; ROC curve

1. Introduction

In many areas of biomedical research, biomarkers can play important roles in classifying a subject or a sample into two or more categories or predicting the subject/sample's probability of being in each category. For example, a key research component in developing a vaccine for an infectious disease is the so-called 'immune correlates study,' which seeks to

^{*}Correspondence to: Fred Hutchinson Cancer Research Center, Public Health Sciences, 1100 Fairview Avenue N., M2-B500, Seattle, WA 98109, USA. yhuang@fhcrc.org.

[#]Equal contribution

identify immune biomarkers that are associated with the risk of infection [1]. Finding immune correlates helps us understand the biological mechanism of vaccine protection and reduces the cost of future vaccine efficacy trials. In practice, a single biomarker often has limited classification/prediction performance, making it desirable to combine multiple biomarkers for better performance.

In this paper we assume there are two outcome categories: case/disease group and control/non-diseased group, denoted by D and $\bar{D}$, respectively. We consider how to combine multiple markers to achieve the best classification performance. Many classification methods have been developed. We will focus on a class of methods which assign a scalar value called score to each subject. Subjects with scores higher than a threshold are then classified as cases, while subjects with lower scores are classified as controls. This is a flexible class of methods since the score can be expressed as a linear or nonlinear combination of the input variables. It does, however, exclude methods such as the artificial neural network and the classic decision tree.

Several performance criteria have been proposed to evaluate classification/prediction procedures. Two basic criteria are sensitivity and specificity. Sensitivity is the probability of correctly identifying a case subject; specificity is the probability of correctly classifying a control subject. Classification error combines the two by weighting 1-sensitivity and 1-specificity according to the probabilities of cases and controls. Both sensitivity and specificity depend on the choice of threshold. When we vary the threshold from $-\infty$ to ∞ , we get pairs of sensitivity and specificity. A plot of sensitivity versus 1-specificity is called the Receiver Operating Characteristic (ROC) curve. See [2] and [3] for a review of the ROC methodology. The area under the ROC curve (AUC) is a commonly used criterion for a model's overall classification performance. It is threshold-independent and equals the probability that a randomly chosen case score is greater than a randomly chosen control score.

Two broad groups of scoring methods have been proposed for AUC-based classification. The first group of methods optimizes a criterion function that can be written as the sum of n terms, where n is the sample size; on the other hand, the criterion function used in the second group of methods is the sum of $O(n^2)$ terms. We refer to the first group as the single-indexed approach and the second group as the double-indexed approach or the AUC approach. Under this classification, both logistic regression, which optimizes a likelihood function that depends on the marker combination, and Support Vector Machine (SVM) [4], which optimizes a hinge loss function, are single-indexed methods. On the other hand, the AUC approach seeks to maximize the empirical AUC, or equivalently, to minimize the empirical AUC loss defined as:

$$\frac{1}{n_D n_{\bar{D}}} \sum_{i=1}^{n_D} \sum_{j=1}^{n_{\bar{D}}} [1 - I\{(x_i - x_j)^T \beta > 0\}] \equiv \frac{1}{n_D n_{\bar{D}}} \sum_{p=1}^{n_D n_{\bar{D}}} \{1 - I(\eta_p^\Delta > 0)\}, \quad (1)$$

where n_D and $n_{\bar{D}}$ are the sample sizes of the case and control group, i and j index the case and control group respectively, $\mathbf{x}_i$ or $\mathbf{x}_j$ is a vector of biomarkers, $\boldsymbol{\beta}$ is a vector of coefficients, p refers to the pair of case indexed by i_p and control indexed by j_p and $\boldsymbol{\eta}_p^\Delta = (\mathbf{x}_{i_p} - \mathbf{x}_{j_p})^T \boldsymbol{\beta}$.

When the training data sample size is large, the empirical AUC approaches the population AUC, and the AUC approach is expected to perform better than the single-indexed approach in general when AUC is chosen as the classification performance criterion [5]. When the training data sample size is small to moderate, the scenario we are most concerned with, the single-indexed approach, may be more efficient, because the data are not reduced to rank as in the AUC approach. However, the AUC approach is more robust to outliers than the single-indexed approach. Methods like logistic regression and SVM are sensitive to outliers because the contribution of each individual observation to the criterion function is unlimited. Even though many proposals have been made to bend the influence of outlying observations to obtain robust methods, e.g. a M-estimator type robust logistic regression method by Bianco and Yohai (1996) [6] and robust statistical learning methods [7, 8, 9], these methods all involve nuisance parameters in their criterion functions, which determines the degree of bending. On the other hand, the robustness of the AUC approach is a result of directly optimizing the criterion function used to evaluate the classification procedure, and there is no need to tune a nuisance parameter related to robustness.

The AUC approach is more challenging than the single-indexed approach in terms of optimization. For example, the logistic likelihood function is smooth and convex, but the empirical AUC function is neither. There have been two main groups of AUC-based methods. The majority of the AUC-based methods are the Smoothed AUC (SAUC) methods. These methods use a smooth sigmoid function $S_s(x)$ to approximate the step function $1 - I(x > 0)$ in the AUC loss (Figure 1a). For example, an approximation based on the normal cumulative distribution function (CDF) Φ , $S_s(x) = \Phi(-x/s)$, was used by [10, 11, 12] among others. Another commonly used approximation is the logistic function $S_s(x) = 1/\{1 + \exp(x/s)\}$, see e.g. [13, 14, 15]. In both normal CDF and logistic approximations, a scale parameter s controls the closeness of the approximations. Smaller s leads to better approximation but harder optimization. Other smooth approximations have also been proposed, e.g. polynomial approximation [16]. While the SAUC criterion functions are differentiable and optimization is relatively easy, the gradient-based methods used in the optimization are difficult to remove from the local optimum and result in sub-optimal solutions.

A second group of methods in the AUC approach is termed the Support Vector Machine-AUC (SVM-AUC) method [17, 18, 19, 20]. These methods replace the step function $1 - I(x > 0)$ in the AUC loss function by a hinge loss function $(1 - x)_+$ (Figure 1b), thus seeking to minimize a convex upper bound of the AUC loss function. The resulting optimization problem is a convex optimization problem and can be solved by support vector machine software. Because the contribution of the difference in a pair of case/control scores can increase infinitely, SVM-AUC does not approximate the AUC loss function well. This problem is particularly serious when the training data is contaminated with outliers.

Most of the AUC-based methods developed so far focus on identifying linear combinations of markers, with the exception of the boosting approach used in [21]. While providing a

simple summary of relationship between markers, a linear combination is limited in classification performance. One way to improve classifier performance is to map the markers to a feature space through basis expansion such as polynomials or splines. Choosing the right expansion is, however, nontrivial and requires substantial amount of experience in data modeling and domain knowledge.

In this paper, we propose a new AUC-maximization method for identifying marker combinations. We start by proposing a new loss function in Section 2, which provides a better approximation to the AUC criterion function than the SVM-AUC methods. We study the asymptotic theory for the estimated linear combination using the proposed method. In Section 3, we develop a difference-of-convex functions algorithm to find the best linear combination. The algorithm avoids some of the pitfalls associated with gradient-based methods used in SAUC methods. We further show that the algorithm lends itself naturally to the application of the so-called ‘kernel trick’ ([22], Chapters 5 and 12), which allows it to find the best linear combination in a higher (often infinite) dimensional feature space without needing to explicitly specify the mapping from the input markers to the features. In Section 4, we present simulation studies to study the comparative performance of the proposed method against common existing methods. In Section 5, we illustrate the application of the proposed method with an immune biomarker dataset from a recent HIV vaccine trial conducted in Thailand. We make concluding remarks in Section 6.

2. The Ramp AUC Estimator

To provide a new approach to finding AUC-based marker combinations that is optimized using a non-gradient based method effectively and can handle nonlinear combinations, we propose to approximate the step function, $1 - \mathbb{I}(x)$, in the empirical AUC loss (1) by a ramp function:

$$h_s(x) = \begin{cases} 1 & \text{if } \frac{x}{s} \leq -\frac{1}{2} \\ -\frac{x}{s} + \frac{1}{2} & \text{if } -\frac{1}{2} \leq \frac{x}{s} < \frac{1}{2} \\ 0 & \text{if } \frac{x}{s} \geq \frac{1}{2} \end{cases},$$

where $s > 0$ is a scale parameter (Figure 1). We call $\sum_{p=1}^{n_D n_{\overline{D}}} h_s(\eta_p^\Delta) / (n_D n_{\overline{D}})$ the Ramp AUC (RAUC) loss function. For any given $\eta_p^\Delta \neq 0$, the value of this loss function moves towards 0 or 1 and becomes 0 or 1 at some point as $s \rightarrow 0$. It is useful to keep track of the proportion of $h_s(\eta_p^\Delta)$ being 0 or 1, termed the saturation ratio, since it reflects the degree of approximation.

There are identifiability issues with all methods in the AUC approach. The problem with minimizing the empirical AUC loss function arises from the fact that $1 - \mathbb{I}(x > 0) = 1 - \mathbb{I}(cx > 0)$, where c is a scalar constant. Minimizing SAUC loss also has an identifiability issue because (i) $S_s(|x|) > S_s(c|x|)$ for $c > 1$ and (ii) $S_s(x) = S_{s-1}(x/s)$ for both normal CDF approximation and logistic approximation. Current SAUC methods use a combination of two procedures to deal with the identifiability problem. First, a constraint is imposed on the

coefficients β . Pepe and Thompson [23] suggested choosing an anchor variable x_1 and setting its coefficient $\beta_1 = 1$, while Lin et al. [24] proposed to constrain the L_1 -norm of β to 1. Second, an empirical rule for selecting s is adopted to set s before optimization.

Gammerman [25] proposed to set $s = \arg\max_s (95\% \eta_p^\Delta / s \geq 5)$ in order to ensure a good approximation of SAUC to the empirical AUC loss function; Vexler et al. [11] proposed to let $s = (n_D n_{\bar{D}})^{-0.1}$ based on simulation studies; and Lin et al. [24] proposed to have $s = n^{-1/3} \text{Stddev}(\mathbf{x} \beta^0 / \|\beta^0\|_2)$, where $n = n_D + n_{\bar{D}}$ and β^0 is an initial estimate of β out of both empirical and asymptotic considerations.

Minimizing RAUC has similar identifiability problems because $h_s(x) = h_s(cx)$ for $c > 1$ and because the ramp function also satisfies $h_s(|x|) = h_{s-1}(|x|/s)$. We propose a new way to address these issues by setting $s = 1$ and penalizing the L_2 -norm of β . That is, we seek to minimize a penalized RAUC loss:

$$\min_{\beta} \sum_{p=1}^{n_D n_{\bar{D}}} h(\eta_p^\Delta) + \frac{1}{2} \lambda_n \|\beta\|_2^2. \quad (2)$$

This effectively minimizes RAUC loss while constraining $\|\beta\|_2^2$ to a constant. But instead of specifying how big $\|\beta\|_2^2$ needs to be, we use λ_n as a lever to adjust it. When λ_n gets smaller, we constrain $\|\beta\|_2^2$ at higher value. Consequently, the saturation ratio increases and the approximation error of the the ramp AUC criterion decreases; but, at the same time, the search algorithm becomes more likely to get stuck in suboptimal solutions. A thorough but computationally intensive approach for selecting λ is to perform searches under a series of λ_n and choose the one that yields the best empirical AUC. On the other hand, a rule of thumb that works well when finding the best linear combinations is that when a choice of λ leads to around 75% saturation ratio, the corresponding solution is often close to being optimal. In our experience, choosing smaller λ_n to get higher saturation ratios frequently results in either negligible improvement or, worse, deterioration in the quality of the solution.

We prefer constraints on L_2 -norm over constraints on an anchor variable coefficient because, as has been pointed out by others [15, 24], there are times when it is difficult to select an anchor variable. Constraining $\|\beta\|_2$ also ensures a more uniform approximation of the AUC loss across the parameter space than constraining $\|\beta\|_1$. Suppose each covariate has been standardized to have mean 0 and standard deviation 1. For both SAUC and RAUC, this means the saturation ratio remains relatively constant for different marker combinations under the $\|\beta\|_2$ constraint but not under the $\|\beta\|_1$ constraint. As the SAUC or RAUC tends to increase as the saturation ratio increases, the estimate of the combination is likely to be more biased in the latter case, particularly when the range of saturation ratios is not close to 1.

Consider the statistical model $\Pr(Y = 1 | X) = G(a + X^T \beta)$, where Y is a binary outcome and G is an unknown monotonically increasing inverse link function. Let the first derivative of $G(\cdot)$ be denoted by $G'(\cdot)$. Let β be d -dimensional and denote $\alpha(\beta) = (\beta_2/\beta_1, \dots, \beta_d/\beta_1)^T$. Denote the true parameters by β_0 and denote $\theta_0 = \alpha(\beta_0)$. Also denote the first element of β_0

by $\beta_{0,1}$ and we assume it not to be 0. Sherman [26] proved maximizing empirical AUC leads to a consistent and asymptotically normal estimate of θ . Ma and Huang [15] showed this was also true for maximizing smoothed AUC. We show here that the estimate of combination $\hat{\theta} = \hat{\alpha}\hat{\beta}$ from minimizing the penalized empirical ramp AUC loss (2) is also consistent and asymptotically normal through two theorems.

Theorem 1 (Consistency)

Under certain regularity conditions and $\lambda_n = o_p(n^2)$, $\hat{\theta} \xrightarrow{p} \theta_0$ as $n \rightarrow \infty$.

Theorem 2 (Asymptotic Normality)

Let $\mathcal{X}$ denote the last $d-1$ components of $\mathbf{X}$, $\eta_0 = \alpha + \mathbf{X}^T \beta_0$, and g_0 denote the marginal density of η_0 . Denote

$$V = \mathbb{E}_{\eta_0} \left\{ \beta_{0,1}^2 \text{Var}(\mathcal{X}|\eta_0) (Y - G(\eta_0))^2 g_0(\eta_0)^2 \right\}$$

$$\Delta = -\mathbb{E}_{\eta_0} \left\{ \beta_{0,1}^2 \text{Var}(\mathcal{X}|\eta_0) G'(\eta_0) g_0(\eta_0) \right\}.$$

Suppose assumptions (A1) – (A4) hold and $\lambda_n = o_p(1)$, then

$$n^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \Delta^{-1} V \Delta^{-1})$$

as $n \rightarrow \infty$.

The assumptions and proofs of the theorems mirror those of [26] and [15] and can be found in Appendix A. Based on the proofs, it is clear that the asymptotic variance of the RAUC estimator of the best linear combination should be the same as that of the empirical AUC estimator and the smoothed AUC estimator. However, our asymptotic variance formula differs from that of [26] and [15] by a factor of 4. The difference can be traced to the expectation of the second derivative of the kernel of the empirical process in Lemma 1 of Appendix A.

To consistently estimate the asymptotic variance, one can either use numerical derivatives as suggested by Sherman [26] in the context of maximizing empirical AUC or use the following empirical smoothed estimate in the spirit of [15]:

$$\hat{V} = \frac{1}{n^3} \hat{\beta}_1^2 \sum_{i=1}^{n_D} \Pi \left(\sum_{j=1}^{n_D} \left[S' \left\{ (\mathbf{x}_i - \mathbf{x}_j)^T \hat{\beta} \right\} (\chi_i - \chi_j) \right] \right) + \frac{1}{n^3} \hat{\beta}_1^2 \sum_{j=1}^{n_D} \Pi \left(\sum_{i=1}^{n_D} \left[S' \left\{ (\mathbf{x}_i - \mathbf{x}_j)^T \hat{\beta} \right\} (\chi_i - \chi_j) \right] \right)$$

$$\hat{\Delta} = \frac{1}{n^2} \hat{\beta}_1^2 \sum_{i=1}^{n_D} \sum_{j=1}^{n_D} \left[S'' \left\{ (\mathbf{x}_i - \mathbf{x}_j)^T \hat{\beta} \right\} (\chi_i - \chi_j) (\chi_i - \chi_j)^T \right],$$

where $\Pi(\mathbf{x}) = \mathbf{x}\mathbf{x}^T$; $S(x)$ is a smooth sigmoid function, e.g. $\Phi(-x)$, with first derivative denoted by $S'(x)$ and second derivative denoted by $S''(x)$; and χ_i is the last $d-1$ components of $\mathbf{x}_i$. As observed by Ma and Huang [15], the performance of the empirical variance

estimate is less than satisfactory when the sample size is small to medium and the more computationally intensive bootstrap methods are preferred.

3. Optimization and Linear and Nonlinear Marker Combination

3.1. Identification of Linear Marker Combinations

The penalized RAUC loss (2) is a non-convex, non-smooth function. To find the best combination, we write the ramp function h as the difference between two convex functions: $h(x) = h_1(x) - h_2(x)$ (Figure 1b)

$$h_1(x) = \left(\frac{1}{2} - x\right)_+, \quad h_2(x) = \left(-\frac{1}{2} - x\right)_+.$$

This allows the application of difference of convex functions algorithm (DCA) [27, 9]. The essence of DCA is to use a series of convex optimization problems to approximate the non-convex problem. This is achieved by iteratively approximating h_2 with a first order Taylor expansion; there is no need to approximate h_1 because the difference between a convex function and a linear function is a convex function. The optimization proceeds as follows:

Step 1. Start with an initial guess for η_p^Δ and assign it to $\eta_p^{\Delta,0}$. For example, we can use a robust logistic regression to get an initial estimate $\hat{\theta}^{rob}$ and let

$$\eta_p^{\Delta,0} \leftarrow (x_{i_p} - x_{j_p})^T \left(1, \hat{\theta}^{rob}\right)^T.$$

Step 2. Solve

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{p=1}^{n_D n_{\overline{D}}} \left\{ h_1(\eta_p^\Delta) - \hat{h}_2(\eta_p^\Delta, \eta_p^{\Delta,0}) + \frac{1}{2} \lambda_n \|\beta\|_2^2 \right\}, \quad (3)$$

where $\hat{h}_2(\eta_p^\Delta, \eta_p^{\Delta,0}) = h_2(\eta_p^{\Delta,0}) + h_2'(\eta_p^{\Delta,0}) \eta_p^\Delta$, and $h_2'(x) = -I(x < -1/2)$ is the first derivative of $h_2(x)$ with respect to x , except 0 at $x = -1/2$.

Step 3. Set $\eta_p^{\Delta,0} \leftarrow (x_{i_p} - x_{j_p})^T \hat{\beta}$ and go back to step 2 until the change in the penalized RAUC loss is less than a pre-specified constant.

In step 2 of the algorithm above, we solve a convex optimization problem. As h_1 and $\hat{h}_2$ are not smooth functions, the standard approach is to convert it to a constrained smooth optimization problem by introducing slack variables ξ_p to replace h_1 and adding two sets of constraints. This leads to:

$$\begin{aligned} \min_{\xi, \beta} \quad & \sum_{p=1}^{n_D n_{\overline{D}}} \left\{ \xi_p - h_2'(\eta_p^{\Delta,0}) \eta_p^\Delta \right\} + \frac{1}{2} \lambda_n \|\beta\|_2^2 \\ \text{subject to} \quad & \xi_p \geq \frac{1}{2} - \eta_p^\Delta, \xi_p \geq 0, \end{aligned} \quad (4)$$

where we have dropped terms void of ξ or β . Because the constraints in the above optimization problem are hard to work with, we seek to find its dual problem [28] via the Lagrange method. The primary Lagrangian is

$$L_p = \sum_{p=1}^{n_D n_{\overline{D}}} \{ \xi_p - h'_2(\eta_p^{\Delta,0}) \eta_p^{\Delta} \} + \frac{1}{2} \lambda_n \|\beta\|_2^2 - \sum_p \alpha_p \left(\xi_p + \eta_p^{\Delta} - \frac{1}{2} \right) - \sum_p \gamma_p \xi_p,$$

where α and γ are non-negative Lagrange multipliers, corresponding to the two sets of constraints in (4). Setting derivatives of the Lagrangian with respect to the primary space variables β and ξ to 0, we arrive at

$$\begin{aligned} \beta &= \frac{1}{\lambda_n} \sum_{p=1}^{n_D n_{\overline{D}}} \{ h'_2(\eta_p^{\Delta,0}) + \alpha_p \} x_p^{\Delta} \\ 1 &= \alpha_p + \gamma_p. \end{aligned} \quad (5)$$

Plugging them back to L_p , the terms having ξ_p vanish. After some algebra (see Appendix B for details), we arrive at the dual problem:

$$\begin{aligned} \min_{\alpha} & \alpha^T Q \alpha - b^T \alpha \\ \text{subject to} & 0 \leq \alpha_p \leq 1 \end{aligned} \quad (6)$$

where Q is a square matrix whose $[p, p']^{th}$ element is $\langle x_p^{\Delta}, x_{p'}^{\Delta} \rangle / \lambda_n$, and $b = 1 - 2 \times Q \times h'_2(\eta^{\Delta,0})$. The optimization problem (6) has a set of simple box constraints and can be solved via many quadratic programming methods. Once we solve the dual problem, solution to the primal problem is found through (5).

Having found the best combination, to classify a new observation we will use the score:

$$\eta_{new} = \sum_{p=1}^{n_D n_{\overline{D}}} \{ h'_2(\eta_p^{\Delta,0}) + \alpha_p \} \langle x_{new}, x_p^{\Delta} \rangle, \quad (7)$$

where $\eta^{\Delta,0}$ and α come from the last iteration of DCA. Similar to the SVM literature [29], we call those x_p^{Δ} with $h'_2(\eta_p^{\Delta,0}) + \alpha_p = 0$ non-support pairs because they do not contribute to the score. There are usually a great number of non-support pairs and this means only some of the inner products $\langle x_{new}, x_p^{\Delta} \rangle$ need to be evaluated, allowing for faster classification.

3.2. Identification of Nonlinear Marker Combinations

As the best classifiers may not be among linear combinations of the input markers, we wish to enlarge the feature space via basis expansion and find the best linear combination in the

enlarged feature space. For example, if we have two markers (x_1, x_2) , we go beyond the linear trends and look at the second order trends as well as the interaction between the two markers. Then we work with the feature space $(x_1, x_2, x_1^2, x_2^2, x_1x_2)$. Let ϕ_i denote the feature vector for subject i . We now show that the algorithm above can be adapted to find the best combination of ϕ_i even if ϕ_i is infinite-dimensional and we don't have explicit knowledge of the mapping from x_i to ϕ_i , as long as we know how to evaluate the inner product $\langle \phi_i, \phi_j \rangle$. In step 1 we let $\eta_p^{\Delta,0} \leftarrow 0$. In step 2, the dual problem that needs to be solved has the same form as (6), where Q is now a square matrix whose $[p, p']^{th}$ element is $\langle \phi_p^\Delta, \phi_{p'}^\Delta \rangle / \lambda_n$. But $\langle \phi_p^\Delta, \phi_{p'}^\Delta \rangle$ can be written as:

$$\langle \phi_p^\Delta, \phi_{p'}^\Delta \rangle = \langle \phi_{i_p}, \phi_{i_{p'}} \rangle + \langle \phi_{j_p}, \phi_{j_{p'}} \rangle - \langle \phi_{i_p}, \phi_{j_{p'}} \rangle - \langle \phi_{j_p}, \phi_{i_{p'}} \rangle.$$

Having solved the dual problem, there is no need to translate it back to the primal space because step 3 calls for $\hat{\eta}$, which has the expression:

$$\hat{\eta}^\Delta = (\phi_1^\Delta, \dots, \phi_p^\Delta) \hat{\beta} = Q \times \{h_2'(\eta^{\Delta,0}) + \hat{\alpha}\}.$$

That is, even though ϕ_i may be infinite-dimensional, $\hat{\eta}$ always has a finite-dimensional representation. We can classify a new observation based on:

$$\begin{aligned} \eta^{new} &= \sum_{p=1}^{n_D n_{\overline{D}}} \{h_2'(\eta_p^{\Delta,0}) + \alpha_p\} \langle \phi(x_{new}), \phi_p^\Delta \rangle \\ &= \sum_{p=1}^{n_D n_{\overline{D}}} \{h_2'(\eta_p^{\Delta,0}) + \alpha_p\} \left\{ \langle \phi(x_{new}), \phi_{i_p} \rangle - \langle \phi(x_{new}), \phi_{j_p} \rangle \right\}, \end{aligned}$$

where again $\eta^{\Delta,0}$ and α come from the last iteration of DCA.

This approach of mapping the input vector to a feature space without having to specify the mapping explicitly has been called the 'kernel trick' in the SVM literature [30, 22]. Let K be a symmetric continuous function that maps $\{x_i\} \times \{x_j\}$ to the real line. We refer to K as a kernel. Mercer's theorem [31, 4] tells us that any positive semidefinite kernel K corresponds to the inner product in some mapping ϕ , i.e. $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$. Thus, if we can compute kernel K , we can find the best linear combination in ϕ . Two commonly used nonlinear kernels include the d^{th} -degree polynomial kernel function $K(x_i, x_j) = (1 + \langle x_i, x_j \rangle)^d$ and the radial basis function (RBF) kernel $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$, where γ is a tuning parameter. A linear kernel $K(x_i, x_j) = \langle x_i, x_j \rangle$ maps a vector to itself, and this allows a uniform treatment of linear and nonlinear combination methods. In order to avoid overfitting in the feature space, we will treat λ_n as a tuning parameter to find the desirable model complexity.

4. Simulation studies

4.1. Linear Marker Combinations

In this simulation study, we examine the performance of RAUC in finding the best linear combination of markers. Two covariates are simulated from a mixture of two bivariate normal random variables

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = (1-\Delta) \times N \left(0, 0.2 \times \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix} \right) + \Delta \times N \left(0, 2 \times \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right),$$

where Δ is a Bernoulli variable with mean π . The outcome variable Y is generated as Bernoulli random variables with means specified by $\text{logit} \{ \Pr (Y = 1) \} = 4X_1 - 3X_2 - (X_1 - X_2)^3$. When no outliers are simulated, π is set to 0; when outliers are simulated, π is set to 0.05. The sample size is 200 for training data and 10^4 for test data. Figure 2 shows two training datasets, one with and one without outliers. The plots suggest that this simulation scenario mimics the the real data example that served as a motivation example in [5].

We compare RAUC with three other classification methods. The logistic method fits a logistic regression model to the training data and uses the fitted model to score the test data. For the SAUC method, we choose normal CDF as the sigmoid function, constrain $\|\beta\|_2 = 1$, and choose the scale parameter as prescribed by Lin et al. [24]. Optimization is carried out by adapting the algorithm from Lin et al. The third method we compare to is the SVM method [32, 33, 4] as implemented by Joachims [34]. We choose SVM for comparison here because it is a commonly used classification method and is also a kernel-based method like the RAUC method. In our categorization of classification methods, SVM is more closely aligned with the likelihood approach because the objective function that it maximizes has the same overall shape as the logistic regression method ([22] Fig 10.4). For all methods, we assume the working model to be a linear combination of x_1 and x_2 .

Summary results from 1000 simulations are shown in Table 1. When there are no outliers in the training data, all methods perform similarly. When there are outliers in the training data, the logistic regression method may break down as illustrated by Figure 2. Table 1 shows that, when the test data does not include outliers, the average test data AUC for the logistic regression method and the SAUC method is 0.639 and 0.667, respectively. Because the AUC measure has a narrow range of going from 0.5 for a useless classification method to 1 for a perfect classification method, we will look at relative increase in effective AUC, defined as $\text{AUC} - 0.5$, when comparing two methods. By this measure, SAUC shows a relative increase in effective AUC of $(0.667 - 0.5)/(0.639 - 0.5) - 1 = 20\%$ over the logistic regression method. By this same measure, RAUC^L, the RAUC method with linear kernel, shows a relative increase in effective AUC of $(0.681 - 0.5)/(0.667 - 0.5) - 1 = 8\%$ over SAUC when the test data contains no outliers. When the test data is simulated with outliers, we see a similar pattern of RAUC outperforming SAUC and SAUC outperforming logistic regression. Since the loss functions minimized by RAUC and SAUC both approximate the AUC loss, the difference in performance between the two can be attributed to the difference in

optimization algorithm and suggests that the difference-of-convex functions algorithm employed by RAUC may be less likely to be stuck in local optima than the gradient-based algorithm used by SAUC. Finally, as we would expect, the performance of SVM^l, the SVM estimator with linear kernel, tracks most closely with the logistic regression method as it is not robust against outliers.

Using the knowledge of logit $\{\Pr(Y = 1)\}$ to classify cases and controls gives us a theoretical AUC that represents an upper bound on the classification performance. From a dataset of 500,000 data points, the theoretical AUC is 0.704 under the ‘no outliers’ scenario and 0.720 under the ‘with outliers’ scenario. Figure A.1 in the Supplementary Materials shows the density curves for the distributions of logit $\{\Pr(Y = 1)\}$ in cases and controls under the two scenarios. Comparing the theoretical AUC to Table 1, we see that a linear combination of X_1 and X_2 can achieve close to theoretical AUC under the ‘no outliers’ scenario but not under the ‘with outliers’ scenario. This makes sense because when there are no outliers the values of X_1 and X_2 are very close to 0 and the higher order term $(X_1 - X_2)^3$ is nearly ignorable. Figure A.2 in the Supplementary Materials plots the joint distributions of $(\hat{\beta}_1, \hat{\beta}_2)$ and the distributions of the ratio $\hat{\beta}_2/\hat{\beta}_1$ from RAUC^l. The figure shows that under finite samples the distribution of $\hat{\beta}_2/\hat{\beta}_1$ behaves well when there are no outliers in the training data but has a long tail when there are outliers in the training data.

We repeat this simulation study and replace the first component of the mixture distribution of (X_1, X_2) with a heavy-tailed bivariate double-exponential distribution using the R package *LaplacesDemon*. Similar results are obtained and the details are described in Supplementary Materials Section D.1.

4.2. Nonlinear Marker Combinations

In this simulation study, we study classifier performance when there is a strong nonlinear pattern in the training data. The radial basis function (RBF) kernel is adopted in RAUC and SVM. We simulate 4 covariates independently from Student’s t distribution with 4 degrees of freedom. When no outliers are simulated, the outcome Y is generated as Bernoulli random variables with means specified as logit $\{\Pr(Y = 1) | X_1, X_2, X_3, X_4\} = 10 \times \{\sin(\pi X_1) + \sin(\pi X_2) + \sin(\pi X_3) + \sin(\pi X_4)\}$. When outliers are simulated, the outcome for samples with $|X_1^2 + X_2^2 + X_3^2 + X_4^2| > 10$ is simulated from a different mean model: $10 \times \{\cos(\pi X_1) + \cos(\pi X_2) + \cos(\pi X_3) + \cos(\pi X_4)\}$. The sample size is 200 for training data and 10^4 for test data.

The use of RBF kernel in RAUC and SVM requires careful consideration of the tuning process. The RBF kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$ contains a scale parameter γ . In addition, the penalty parameter λ takes on a more important role as it controls not only approximation error but also model complexity because the implicit feature vector associated with RBF kernel is infinite-dimensional. We perform a grid search on (λ, γ) . The grid points for γ are chosen to be the 50%, 75% and 90% quantiles of the between class Euclidean distance in the training data [35]. The grid points for λ are chosen more adaptively in the following way. Assuming the markers are all scaled to have variance 1, we start with a default grid of seven values for λ : 2^{-3} , 2^{-2} , 2^{-1} , 1, 2, 2^2 , and 2^3 . We evaluate the

performance of the classifiers at the 3×7 lattice. If the optimal $(\hat{\lambda}, \hat{\gamma})$ has a $\hat{\lambda}$ that corresponds to either the minimum or maximum tested λ , we expand the lattice to include a new λ^* that is another two-fold away from 1 than $\hat{\lambda}$. We repeat this process until the optimal λ is not the minimum or maximum tested λ . To evaluate tuning parameters, we generate a tuning dataset of size 10^3 . The mean and standard error of the test data AUC over 1000 replications are shown in Table 2.

When the training data is free of outliers, both logistic regression method and SAUC method do not perform well due to strong nonlinearity in the data. Both RAUC method and SVM method with RBF kernel perform satisfactorily. As expected, the test data AUC is lower when the test data contains outliers than when the test data does not contain outliers.

When the training data contains outliers, we also observe that RAUC and SVM with RBF kernel outperform logistic and SAUC. Furthermore, RAUC with RBF kernel also shows an advantage over SVM with RBF kernel. The relative increase in effective AUC is $(0.817 - 0.5)/(0.790 - 0.5) - 1 = 9\%$ when the test data has no outliers, and $(0.816 - 0.5)/(0.795 - 0.5) - 1 = 7\%$ when the test data also has outliers.

From a dataset of 500,000 data points, the theoretical AUC is 0.995 under both ‘no outliers’ and ‘with outliers’ scenarios. Figure B.1 in the Supplementary Materials shows the density curves for the distributions of $\text{logit}\{\Pr(Y=1)\}$ in cases and controls under the two scenarios. Figure B.2 in the Supplementary Materials shows scatterplots of ‘theoretical combination’, i.e. $\text{logit}\{\Pr(Y=1)\}$, and RAUC^r estimated combination for the training data under the two scenarios from one Monte Carlo replicate. The plots show that for the bulk of the 200 data points the correlation between theoretical and estimated combinations are high, but when the theoretical combination is near 0, the estimated combination may show wide variability.

We repeat this simulation study with (X_1, X_2, X_3, X_4) simulated from a correlated multivariate Student’s t distribution, and obtain similar results (details in Supplementary Materials Section D.2).

5. Data example

RV144 is a community-based, randomized, multicenter, double-blind, placebo-controlled HIV vaccine efficacy trial conducted in Thailand from 2003 to 2006 [36]. In the modified intention-to-treat analysis involving 16,395 subjects, the vaccine efficacy was 31.2% with a P-value of 0.04. In an effort to identify immune correlates of infection risk potentially relevant for vaccine-induced protection, Haynes et al. [1] conducted a case-control study using biomarker measurement from blood samples of 41 cases and 205 controls, all from the vaccine arm. Haynes et al. reported an array of immunological response variables measured from these samples. We choose two of them for illustration: *IgA*, which measures the IgA class of serum antibody binding to HIV envelope protein gp120, and *IL13*, which measures the amount of *IL13* production by peripheral blood mononuclear cells when stimulated by gp120 peptide. Both variables are continuous variables and are scaled to have standard deviation 1. We perform repeated random sub-sampling validation for AUC estimation. In

iteration, we randomly split the data into 4/5 training set and 1/5 test set stratified on the case status. Test data AUC over 1000 splits is reported in Table 3.

For methods using the RBF kernel, we need a way to tune the tuning parameters. Because the number of cases is limited, it does not work well to further split the training set into a training subset and a tuning subset. Instead, we adopt the following strategy. We will use the same tuning parameters value for all training/test datasets. To find a good value for the tuning parameters, we randomly split the data into 4/5 training set and 1/5 tuning set, and perform a grid search to maximize the classification performance on the tuning data. The median value from 1000 such splits is chosen as the tuning parameter to be used in the training/test phase. For SVM, we try both this strategy and generalized approximate cross validation (GACV [37]) and find that our strategy performs slightly better than GACV; hence, we only report results from the former strategy.

The results in Table 3 show that, among the methods looking for the best linear combination of markers, the performance of RAUC with linear kernel is on par with other methods. The RAUC method detects a nonlinear trend in the relationship between infection and the two markers. Comparing $RAUC^I$ and $RAUC^J$, the relative increase in effective AUC is $(0.656 - 0.5)/(0.621 - 0.5) - 1 = 29\%$. To help interpret the nonlinearity, we make scatterplots of IgA and $IL13$ in Figure A.1 of Supplementary Materials, where cases are represented by red circles, controls are represented by black triangles, and classifier decision boundaries (left: $RAUC^I$ and right: $RAUC^J$) are represented by shaded or non-shaded background. For $RAUC^I$, $IL13$ is drawn on the linear scale; for $RAUC^J$, $IL13$ is drawn on a transformed scale to provide a clearer view. The shape of the decision boundary from $RAUC^J$ suggests that the nonlinearity can be summarized by a simple AND/OR rule: $IgA < 1$ AND $IL13 > c_1$ OR $IgA > 1$ AND $IL13 > c_1 + \delta$. The SVM method with nonlinear kernel does not perform as well, which is likely the manifestation of small sample performance issue with SVM^r.

6. Concluding Remarks

In this paper we propose a new criterion function, the penalized RAUC loss function, for finding marker combinations with optimal classification/prediction performance. By using different kernels, the RAUC method can find both linear and nonlinear best marker combinations. We show that the best linear marker combination is a consistent and asymptotically normal estimator of the true marker combination when the data is generated from a semiparametric generalized linear model.

In the numerical studies, we show that the RAUC method with linear kernel out-performs SAUC as a linear classification method under some scenarios. While these two methods are both consistent estimators of the true marker combination and have the identical asymptotic variance, the difference-of-convex functions algorithm employed by RAUC is less likely to be stuck in local optima than the gradient-based algorithm used by SAUC. In the presence of nonlinear patterns, the RAUC method with nonlinear kernels allows a flexible way to derive nonlinear marker combinations for optimizing classification accuracy, with performance better or comparable to common alternative estimators. R package for implementing the RAUC method is available from the authors upon request.

Our focus in this paper is to develop classification rules to flexibly combine several candidate biomarkers that have been selected by investigators as potentially useful. Another important aspect of biomarker studies is to use data to help select important markers and features. Variable selection is well developed in regression analysis with linear predictors, where many types of penalties have been proposed and oracle properties established [38]. In AUC analysis with linear marker combinations, the penalization approach to variable selection faces an added layer of complexity due to the fact that a constraint on β is needed to ensure identifiability. Some efforts have been made in this area, e.g. Lin et al. [24] proposed to use a SCAD penalty [39] while maintaining $\|\beta\| = 1$. In AUC analysis with nonlinear marker combinations, model selection is further complicated by the fact that the input markers are mapped to a higher-dimensional feature space and parsimony may be sought on the feature level or on the marker level. Feature-level model selection has been studied in nonlinear kernel SVM [40, 41] and marker-level model selection has been explored in Gaussian process regression models [42, 43]. Our framework can be naturally extended to perform marker-level model selection for AUC analysis and is currently under investigation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Krisztian Sebestyen for programming assistance. We also thank the Editor, the AE and the two referees for their insightful comments that helped improve the manuscript. This work was supported by the cooperative agreement W81XWH-07-2-0067 between the Henry M. Jackson Foundation and the Department of Defense, and National Institute of Allergy and Infectious Diseases grants 1R56AI116369-01A1, R01-GM106177 and UM1AI068635.

References

1. Haynes BF, Gilbert PB, McElrath MJ, Zolla-Pazner S, Tomaras GD, Alam SM, Evans DT, Montefiori DC, Karnasuta C, Sutthent R, et al. Immune-correlates analysis of an HIV-1 vaccine efficacy trial. *New England Journal of Medicine*. 2012; 366(14):1275–1286. [PubMed: 22475592]
2. Pepe, M. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press; 2003. Oxford statistical science series <http://books.google.com/books?id=kMyXEJEtFmkC>
3. Zhou, X.; Obuchowski, N.; McClish, D. *Statistical Methods in Diagnostic Medicine*. Wiley-Interscience; 2002. Wiley series in probability and statistics http://books.google.com/books?id=ijN_Dlx7wmoC
4. Vapnik, V. *The Nature of Statistical Learning Theory*. Springer Verlag; 1995.
5. Pepe M, Cai T, Longton G. Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics*. 2006; 62(1):221–229. [PubMed: 16542249]
6. Bianco A, Yohai V. Robust estimation in the logistic regression model. *Robust Statistics, Data Analysis, and Computer Intensive Methods, Lecture Notes in Statistics*. 1996; 109:17–34.
7. Wu Y, Liu Y. Robust truncated hinge loss support vector machines. *Journal of the American Statistical Association*. 2007; 102(479):974–983.
8. Liu Y, Shen X. Multicategory ψ -learning. *Journal of the American Statistical Association*. 2006; 101(474):500–509.
9. Liu Y, Shen X, Doss H. Multicategory ψ -learning and support vector machine: computational tools. *Journal of Computational and Graphical Statistics*. 2005; 14(1):219–236.

10. Lloyd C. Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. *Journal of the American Statistical Association*. 1998; 93(444):1356–1364.
11. Vexler A, Liu A, Schisterman E, Wu C. Note on distribution-free estimation of maximum linear separation of two multivariate distributions. *Journal of Nonparametric Statistics*. 2006; 18(2):145–158.
12. Zhou X, Chen B, Xie Y, Tian F, Liu H, Liang X. Variable selection using the optimal ROC curve: An application to a traditional chinese medicine study on osteoporosis disease. *Statistics in Medicine*. 2012; 31(7):628–635. [PubMed: 21290404]
13. Yan, L.; Dodier, R.; Mozer, M.; Wolniewicz, R. Optimizing classifier performance via an approximation to the Wilcoxon-Mann-Whitney statistic. *Proceedings of the Twentieth International Conference on Machine Learning; The AAAI Press; 2003*.
14. Herschtal, A.; Raskutti, B. Optimising area under the ROC curve using gradient descent. *Proceedings of the Twenty-First International Conference on Machine Learning; ACM; 2004*.
15. Ma S, Huang J. Combining multiple markers for classification using ROC. *Biometrics*. 2007; 63(3):751–891533. [PubMed: 17825007]
16. Calders T, Jaroszewicz S. Efficient auc optimization for classification. *Knowledge Discovery in Databases: PKDD 2007*. 2007:42–53.
17. Brefeld, U.; Scheffer, T. *Proceedings of the ICML 2005 Workshop on ROC Analysis in Machine Learning*. 2005. AUC maximizing support vector learning.
18. Rakotomamonjy A. Optimizing area under roc curve with svms. *Proceedings of ECAI 04 ROC and Artificial Intelligence Workshop*. 2004
19. Joachims, T. A support vector method for multivariate performance measures. *Proceedings of the 22nd International Conference on Machine Learning; ACM; 2005*. p. 377–384.
20. Wang Y, Chen H, Li R, Duan N, Lewis-Fernández R. Prediction-based structured variable selection through the receiver operating characteristic curves. *Biometrics*. 2011; 67(3):896–905. [PubMed: 21175555]
21. Komori O, Eguchi S. A boosting method for maximizing the partial area under the roc curve. *BMC Bioinformatics*. 2010; 11(1):314. [PubMed: 20537139]
22. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2. Springer; 2009. Springer Series in Statistics
23. Pepe M, Thompson M. Combining diagnostic test results to increase accuracy. *Biostatistics*. 2000; 1(2):123–140. [PubMed: 12933515]
24. Lin H, Zhou L, Peng H, Zhou X. Selection and combination of biomarkers using roc method for disease classification and prediction. *Canadian Journal of Statistics*. 2011; 39(2):324–343.
25. Gammerman, A. *Computational Learning and Probabilistic Reasoning*. John Wiley & Sons, Inc; 1996.
26. Sherman R. The limiting distribution of the maximum rank correlation estimator. *Econometrica: Journal of the Econometric Society*. 1993; 61(1):123–137.
27. An LTH, Tao PD. Solving a class of linearly constrained indefinite quadratic problems by dc algorithms. *Journal of Global Optimization*. 1997; 11(3):253–285.
28. Boyd, S.; Vandenberghe, L. *Convex Optimization*. Cambridge University Press; 2004. <http://books.google.com/books?id=mYm0bLd3fcoC>
29. Burges C. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*. 1998; 2(2):121–167.
30. Aizerman A, Braverman E, Rozoner L. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*. 1964; 25:821–837.
31. Mercer J. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London Series A, Containing Papers of a Mathematical or Physical Character*. 1909; 209:415–446.
32. Boser, B.; Guyon, I.; Vapnik, V. A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory; ACM; 1992*. p. 144–152.
33. Cortes C, Vapnik V. Support-vector networks. *Machine learning*. 1995; 20(3):273–297.

34. Joachims, T. Learning to Classify Text Using Support Vector Machines. Vol. 668. Springer; Netherlands: 2002.
35. Brown M, Grundy W, Lin D, Cristianini N, Sugnet C, Furey T, Ares M, Haussler D. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*. 2000; 97(1):262.
36. Rerks-Ngarm S, Pitisuttithum P, Nitayaphan S, Kaewkungwal J, Chiu J, Paris R, Premisri N, Namwat C, de Souza M, Adams E, et al. Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. *New England Journal of Medicine*. 2009; 361(23):2209–2220. [PubMed: 19843557]
37. Wahba, G.; Lin, Y.; Zhang, H. Margin-like quantities and generalized approximate cross validation for support vector machines. *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop*; IEEE; 1999. p. 12–20.
38. Bühlmann, P.; van de Geer, S. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer; Berlin Heidelberg: 2011. Springer Series in Statistics <https://books.google.com/books?id=S6jYXmh988UC>
39. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*. 2001; 96(456):1348–1360.
40. Fung G, Mangasarian O. A feature selection newton method for support vector machine classification. *Computational Optimization and Applications*. 2004; 28(2):185–202.
41. Mangasarian, OL.; Kou, G. Feature selection for nonlinear kernel support vector machines. *Data Mining Workshops, 2007. ICDM Workshops 2007. Seventh IEEE International Conference on*; IEEE; 2007. p. 231–236.
42. Ni X, Zhang D, Zhang H. Variable selection for semiparametric mixed models in longitudinal studies. *Biometrics*. 2010; 66(1):79–88. [PubMed: 19397585]
43. Savitsky T, Vannucci M, Sha N. Variable selection for nonparametric gaussian process priors: Models and computational strategies. *Statistical Science*. 2011; 26(1):130–149. [PubMed: 24089585]
44. Han A. Non-parametric analysis of a generalized regression model* 1:: The maximum rank correlation estimator. *Journal of Econometrics*. 1987; 35(2–3):303–316.
45. Nolan D, Pollard D. u -processes: Rates of convergence. *The Annals of Statistics*. 1987; 15(2):780–799.

Appendix

Appendix A: Proofs of Theorems 1 and 2

For proof of the consistency, we need two assumptions. (A1) θ_0 is an interior point of Θ which is a compact subset of $\mathbb{R}^{d-1}$. (A2) The support of the covariate vector $\mathbf{x}$ is not contained in any proper linear subspace of $\mathbb{R}^d$, and the first component of $\mathbf{x}$ has a everywhere positive Lebesgue density, conditional on the other components.

Under (A1) and (A2), Han (1987) [44] proved the consistency of the empirical AUC

estimator. Denote the AUC loss by $AUC_n(\theta) = \frac{1}{n^2} \sum_{p=1}^{n_D n_D} I(\eta_p^\Delta < 0)$. It suffices to show that $\sup_{\theta \in \Theta} |AUC_n(\theta) - \argmin_{\beta_1} L_n(\beta)| \rightarrow 0$ in probability as $\lambda_n \rightarrow 0$, where

$$L_n = \frac{1}{n^2} \sum_{p=1}^{n_D n_D} r \left(\eta_p^\Delta \right) + \frac{1}{2n^2} \lambda_n \|\beta\|_2^2. \text{ But}$$

$$\begin{aligned}
\left| \text{AUC}_n(\theta) - \arg\min_{\beta_1} L_n(\beta) \right| &= \left| \frac{1}{n^2} \sum_{p=1}^{n_D n_{\overline{D}}} I(\eta_p^\Delta < 0) - \frac{1}{n^2} \sum_{p=1}^{n_D n_{\overline{D}}} r(\eta_p^\Delta) - \frac{1}{2n^2} \lambda_n \hat{\beta}_1^2 \|\theta\|_2^2 \right| \\
&\leq \frac{1}{n^2} \sum_{p=1}^{n_D n_{\overline{D}}} |I(\eta_p^\Delta < 0) - r(\eta_p^\Delta)| + (\lambda_n/n^2)^{1/2} O_p(1) \\
&\leq \frac{1}{n^2} \sum_{p=1}^{n_D n_{\overline{D}}} I \left[(1, \theta) (x_{i_p} - x_{j_p}) < (\lambda_n/n^2)^{1/4} O_p(1) \right] + (\lambda_n/n^2)^{1/2} O_p(1)
\end{aligned}$$

As argued in [15], the first term in the last line converges to 0 uniformly in probability when $\lambda_n/n^2 \rightarrow 0$. Thus ends the proof of the consistency.

Let F and f be the cdf and pdf for $T = (X_i - X_j)^T \beta$, where X_i is a random sample from the case and X_j is a random sample from the control. Denote the pdf for $[1, \theta] (X_i - X_j)$ by ψ .

$$\begin{aligned}
\mathbb{E} \{r(T)\} &= \Pr(T < -0.5) + \int_{-0.5}^{0.5} \left(\frac{1}{2} - T\right) f(T) d\eta^\Delta \\
&= \frac{1}{2} \{F(0.5) + F(-0.5)\} - \int_{-0.5}^{0.5} T f(T) d\eta^\Delta.
\end{aligned}$$

Take the derivative with respect to β_1 , and as $\beta_1 \rightarrow \infty$,

$$\begin{aligned}
\frac{\partial}{\partial \beta_1} \mathbb{E} \{r(T)\} &= \int_0^{0.5/\beta_1} T \{\psi(-T) - \psi(T)\} d\eta^\Delta \\
&= \frac{0.5}{\beta_1} \times \frac{0.5}{\beta_1} \left\{ \psi\left(-\frac{0.5}{2\beta_1}\right) - \psi\left(\frac{0.5}{2\beta_1}\right) \right\} + o_p(\beta_1^{-3}) \\
&= \beta_1^{-3} \left\{ -\frac{1}{8} \psi'(0) + o_p(1) \right\}.
\end{aligned}$$

Hence, $\mathbb{E} \{r(T)\} + \frac{1}{2n^2} \lambda_n \|\beta\|_2^2$ as a function of β_1 is minimized at

$$\begin{aligned}
0 &= \beta_1^{-3} \left\{ -\frac{1}{8} \psi'(0) + o_p(1) \right\} + \frac{\lambda_n}{n^2} \beta_1 \|\theta\|^2 \\
\beta_1 &= \left(\frac{\lambda_n}{n^2} \right)^{-1/4} \left\{ \frac{1}{8\|\theta\|^2} \psi'(0) + o_p(1) \right\}^{1/4}
\end{aligned}$$

Denote $s_n = \left(\frac{\lambda_n}{n^2} \right)^{1/4} \left\{ \left(8\|\theta_0\|^2 \right)^{-1} \psi'(0) \right\}^{-1/4}$. When $\lambda_n = o_p(1)$, $s_n = o_p(1/\sqrt{n})$. Define $r_n(x) = r(x/s_n)$ and $H_n(\theta) = \frac{1}{n^2} \sum_{i \in D} \sum_{j \in \overline{D}} r_n \{ [1, \theta] (x_i - x_j) \}$.

Let $Z = (Y, X)$. A key term in the normality proof is

$$\tau_n(Z_1, \theta) = \mathbb{E}_{Y_2} [I(Y_1 > Y_2) r_n \{ (1, \theta) (X_1 - X_2) \}] + \mathbb{E}_{Y_2} [I(Y_2 > Y_1) r_n \{ (1, \theta) (X_2 - X_1) \}].$$

By $r_n(x) = 1 - r_n(-x)$, $\tau_n(Z_1, \theta)$ can be simplified to

$$\begin{aligned}
\tau_n(Z_1, \theta) &= \int I(Y_1 > Y_2) r_n \{(1, \theta)(X_1 - X_2)\} dF(X_2, Y_2) + \int I(Y_2 > Y_1) [1 - r_n \{(1, \theta)(X_1 - X_2)\}] dF(X_2, Y_2) \\
&= \int \{I(Y_1 > Y_2) - I(Y_2 > Y_1)\} r_n \{(1, \theta)(X_1 - X_2)\} dF(X_2, Y_2) + \int I(Y_2 > Y_1) dF(X_2, Y_2) \\
&= \int S(Y_1, \alpha + X_2^T \beta_0) r_n \{(1, \theta)(X_1 - X_2)\} dF(X_2) + \int I(Y_1 > Y_2) dF(X_2, Y_2),
\end{aligned}$$

where in the last line, we introduce $S(y, t)$,

$$\begin{aligned}
S(y, t) &= E_{Y|\alpha+X^T\beta_0=t} \{I(y > Y) - I(y < Y)\} \\
&= \begin{cases} \Pr(Y=0|\alpha+X^T\beta_0=t) & \text{if } y=1 \\ -\Pr(Y=1|\alpha+X^T\beta_0=t) & \text{if } y=0 \end{cases} \\
&= y - G(t).
\end{aligned}$$

We now state the third assumption needed for the normality. (A3) (i) and V exist and is negative definite; (ii) partial derivative of g_0 exists and is bounded on the support of (Y, X) ; (iii) for each z , all mixed second partial derivative of $\tau_n(z, \cdot)$ exist on a neighborhood of θ_0 ; (iv) There is an integrable function $M(z)$ such that for all z and θ in a neighborhood of θ_0 , then $\|\nabla_2 \tau_n(z, \theta) - \nabla_2 \tau_n(z, \theta_0)\| < M(z)|\theta|$. We now state and prove a lemma needed for normality.

Lemma 1

Under assumption (A3),

$$\nabla_1 \tau_n(Z, \theta_0) = -\beta_{0,1} (\mathcal{X} - \mathcal{X}_0) S(Y, \eta_0) g_0(\eta_0) + O_p(s_n) \quad (8)$$

and

$$\mathbb{E} [\nabla_1 \tau_n(Z, \theta_0) \{\nabla_1 \tau_n(Z, \theta_0)\}^T] = V + O_p(s_n) \quad (9)$$

$$\mathbb{E} \{\nabla_2 \tau_n(Z, \theta_0)\} = 2\Delta + O_p(s_n). \quad (10)$$

Proof

Let u_i denote the unit vector in $\mathbb{R}^{d-1}$ with i^{th} component equal to unity and ∇^i for the i^{th} component of ∇_1 . By definition of derivative, we have

$$\nabla_i^1 \tau_n(Z, \theta_0) = \lim_{\varepsilon \rightarrow 0} \varepsilon^{-1} [\tau_n(Z, \theta_0 + \varepsilon u_i) - \tau_n(Z, \theta_0)].$$

By the definition of r_n , we can compute

$$\begin{aligned}
& \tau_n(Z, \theta_0 + \varepsilon u_i) - \tau_n(Z, \theta_0) = \int [r_n \{(1, \theta_0 + \varepsilon u_i)(X_1 - X_2)\} - r_n \{(1 - \theta_0)(X_1 - X_2)\}] S(Y_1, \alpha + X_2^T \beta_0) dF(X_2) \\
& = \int \left[\begin{aligned} & I \{(1, \theta_0)(X_1 - X_2) + \varepsilon(X_{1,i} - X_{2,i}) < -s_n/2\} \\ & - I \{(1, \theta_0)(X_1 - X_2) < -s_n/2\} \\ & + I \{|(1, \theta_0)(X_1 - X_2) + \varepsilon(X_{1,i} - X_{2,i})| < s_n/2\} [\tfrac{1}{2} - \{(1, \theta_0)(X_1 - X_2) + \varepsilon(X_{1,i} - X_{2,i})\}] \\ & - I \{|(1, \theta_0)(X_1 - X_2)| < s_n/2\} [\tfrac{1}{2} - (1, \theta_0)(X_1 - X_2)] \end{aligned} \right] S(Y_1, \alpha + X_2^T \beta_0) dF(X_2) \\
& = \int \left[\begin{aligned} & I \{-\alpha + (X_1 - X_2)^T \beta_0 + \varepsilon \beta_{0,1}(X_{1,i} - X_{2,i}) < -\alpha - \beta_{0,1}s_n/2\} \\ & - I \{-\alpha + (X_1 - X_2)^T \beta_0 < -\alpha - \beta_{0,1}s_n/2\} \\ & + I \left\{ \left| (X_1 - X_2)^T \beta_0 + \varepsilon \beta_{0,1}(X_{1,i} - X_{2,i}) \right| < |\beta_{0,1}|s_n/2 \right\} [\tfrac{1}{2} - \{(1, \theta_0)(X_1 - X_2) + \varepsilon(X_{1,i} - X_{2,i})\}] \\ & - I \left\{ \left| (X_1 - X_2)^T \beta_0 \right| < |\beta_{0,1}|s_n/2 \right\} [\tfrac{1}{2} - (1, \theta_0)(X_1 - X_2)] \end{aligned} \right] S(Y_1, \alpha + X_2^T \beta_0) dF(X_2) \\
& = \int \left[\begin{aligned} & \text{let } t = \alpha + X_2^T \beta_0 \\ & - I \{\alpha + \beta_{0,1}s_n/2 + X_1^T \beta_0 < t < \alpha + \beta_{0,1}s_n/2 + X_1^T \beta_0 + \varepsilon \beta_{0,1}(X_{1,i} - X_{2,i})\} \\ & + I \{|t - \alpha + \varepsilon \beta_{0,1}(X_{1,i} - X_{2,i})| < |\beta_{0,1}|s_n/2\} [\tfrac{1}{2} - \{(t - \alpha)/\beta_{0,1} + \varepsilon(X_{1,i} - X_{2,i})\}] \\ & - I \{|t - \alpha| < |\beta_{0,1}|s_n/2\} [\tfrac{1}{2} - (t - \alpha)/\beta_{0,1}] \end{aligned} \right] S(Y_1, t) g_0(t | \mathcal{X}_2) dt dF(\mathcal{X}_2) \\
& = -\varepsilon \beta_{0,1} \int (X_{1,i} - X_{2,i}) S(Y_1, \alpha + X_1^T \beta_0) g_0(\alpha + X_1^T \beta_0 | \mathcal{X}_2) dF(\mathcal{X}_2) + \varepsilon O_p(s_n) + o_p(|\varepsilon|)
\end{aligned}$$

(8) and (9) follows.

To show (10),

$$\begin{aligned}
\nabla_i^2 \tau_n(Z, \theta_0) &= -\beta_{0,1} \int (X_{1,i} - X_{2,i})^2 G'(\alpha + X_1^T \beta_0) g_0(\alpha + X_1^T \beta_0 | \mathcal{X}_2) dF(\mathcal{X}_2) \quad \text{b/c } E\{S(Y, t) | X' \beta_0 = t\} = 0 \\
&= -\beta_{0,1} G'(\alpha + X_1^T \beta_0) \int (X_{1,i} - X_{2,i})^2 p(\eta_2 = \alpha + X_1^T \beta_0 | \mathcal{X}_2) p(\mathcal{X}_2) d\mathcal{X}_2 \\
&= -\beta_{0,1} G'(\alpha + X_1^T \beta_0) \int (X_{1,i} - X_{2,i})^2 g(\alpha + X_1^T \beta_0) p(\mathcal{X}_2 | \eta_2 = \alpha + X_1^T \beta_0) d\mathcal{X}_2 \\
&= -\beta_{0,1} G'(\alpha + X_1^T \beta_0) g(\alpha + X_1^T \beta_0) \int (X_{1,i}^2 - 2X_{1,i}X_{2,i} + X_{2,i}^2) p(\mathcal{X}_2 | \alpha + X_1^T \beta_0) d\mathcal{X}_2 \\
&= -\beta_{0,1} G'(\alpha + X_1^T \beta_0) g(\alpha + X_1^T \beta_0) \{X_{1,i}^2 - 2X_{1,i}E(X_{2,i} | \alpha + X_1^T \beta_0) + E(X_{2,i}^2 | \alpha + X_1^T \beta_0)\} \\
&= -\beta_{0,1} G'(\alpha + X_1^T \beta_0) g(\alpha + X_1^T \beta_0) \left[\{X_{1,i} - E(X_{2,i} | \alpha + X_1^T \beta_0)\}^2 + E(X_{2,i}^2 | \alpha + X_1^T \beta_0) - E^2(X_{2,i} | \alpha + X_1^T \beta_0) \right] \\
&= -\beta_{0,1} G'(\alpha + X_1^T \beta_0) g(\alpha + X_1^T \beta_0) \left[\{X_{1,i} - E(X_{1,i} | \alpha + X_1^T \beta_0)\}^2 + V(X_{1,i} | \alpha + X_1^T \beta_0) \right]
\end{aligned}$$

The rest of the normality proof tracks the proof of Theorem 4 in [26] closely. Denote

$$\Gamma_n(\theta) = H_n(\theta) - H_n(\theta_0) + \frac{1}{2n^2} \lambda_n \hat{\beta}_1^2 \|\theta - \theta_0\|_2^2. \text{ We will show that}$$

$$\Gamma_n(\theta) = \frac{1}{2}(\theta - \theta_0)' \Delta(\theta - \theta_0) + \frac{1}{\sqrt{n}}(\theta - \theta_0)' W_n + o_p(|\theta - \theta_0|^2) + O_p(s_n |\theta - \theta_0|) + o_p\left(\frac{1}{n}\right) \quad (11)$$

uniformly in $\alpha(1)$ neighborhoods of θ_0 , where W_n converges in distribution to a $N(0, V)$ random vector. By Theorem 1 and 2 of [26] and by the fact that $s_n = o_p(1/\sqrt{n})$, the result follows.

For each (z_1, z_2) in $S \otimes S$ and each θ in Θ , define $f_n(z_1, z_2, \theta) = I(y_1 > y_2)[r_n\{(1, \theta)(x_1 - x_2)\} - r_n\{(1, \theta_0)(x_1 - x_2)\}]$. As in [26], let P_n be the empirical measure and U_n be the U-process operator while $Pf_n(u, \cdot, \theta)$ denotes the conditional expectation of $f_n(u, v, \theta)$ given its first argument and $Pf_n(\cdot, v, \theta)$ the conditional expectation of $f_n(u, v, \theta)$ given its second argument. Then we can write $H_n(\theta) - H_n(\theta_0) = \Gamma_n(\theta) + P_n g_n(\cdot, \theta) + U_n r_n(\cdot, \cdot, \theta)$, where

$$\begin{aligned}\Gamma_{n0}(\theta) &= E(f_n(Z_1, Z_2, \theta)) \\ g_n(z, \theta) &= P f_n(z, \cdot, \theta) + P f_n(\cdot, z, \theta) - 2\Gamma_{n0}(\theta) \\ r_n(z_1, z_2, \theta) &= f_n(z_1, z_2, \theta) - P f_n(z_1, \cdot, \theta) - P f_n(\cdot, z_2, \theta) + \Gamma_{n0}(\theta).\end{aligned}$$

We now show that

$$\Gamma_{n0}(\theta) = \frac{1}{2}(\theta - \theta_0)' \Delta(\theta - \theta_0) + o_p(|\theta - \theta_0|^2) + O_p(s_n|\theta - \theta_0|) \quad (12)$$

Note that $2\Gamma_{n0}(\theta) = E(\tau_n(z, \theta) - \tau_n(z, \theta_0))$. Expand $\tau_n(z, \theta)$ about θ_0 to get

$$\tau_n(z, \theta) - \tau_n(z, \theta_0) = (\theta - \theta_0)' \nabla_1 \tau_n(z, \theta_0) + \frac{1}{2}(\theta - \theta_0)' \nabla_2 \tau_n(z, \theta^*)(\theta - \theta_0),$$

where θ^* between θ and θ_0 . By (A3)(iv), for each z in S and each θ in a neighborhood of θ_0 ,

$$\|(\theta - \theta_0)' [\nabla_2 \tau_n(z, \theta) - \nabla_2 \tau_n(z, \theta_0)](\theta - \theta_0)\| < M(z)|\theta - \theta_0|^3.$$

Thus by Lemma 1,

$$\begin{aligned}2\Gamma_{n0}(\theta) &= E\{\tau_n(z, \theta) - \tau_n(z, \theta_0)\} \\ &= (\theta - \theta_0)' E\{\nabla_1 \tau_n(z, \theta_0)\} + E\left\{\frac{1}{2}(\theta - \theta_0)' \nabla_2 \tau_n(z, \theta^*)(\theta - \theta_0)\right\} \\ &= O_p(s_n|\theta - \theta_0|) + (\theta - \theta_0)' E\{(\mathcal{X} - \mathcal{X}_0) S(Y, \eta_0) g_0(\eta_0)\} + (\theta - \theta_0)' \Delta(\theta - \theta_0) + o(|\theta - \theta_0|^2) \\ &= O_p(s_n|\theta - \theta_0|) + (\theta - \theta_0)' \Delta(\theta - \theta_0) + o(|\theta - \theta_0|^2).\end{aligned}$$

Next we show that

$$P_n g_n(\cdot, \theta) = \frac{1}{\sqrt{n}}(\theta - \theta_0)' W_n + O(s_n|\theta - \theta_0|) + o(|\theta - \theta_0|^2) \quad (13)$$

uniformly over $\alpha(1)$ neighborhood of θ_0 , where W_n converges in distribution to a $N(0, V)$ random vector. Note that

$$g_n(z, \theta) = \tau(z, \theta) - \tau(z, \theta_0) - 2\Gamma_{n0}(\theta).$$

Then we have

$$\begin{aligned}
P_n(g_n(z, \theta)) &= P_n(\tau_n(z, \theta) - \tau_n(z, \theta_0)) - 2\Gamma_{n0}(\theta) \\
&= P_n \left((\theta - \theta_0)' \nabla_1 \tau_n(z, \theta_0) + \frac{1}{2} (\theta - \theta_0)' \nabla_2 \tau_n(z, \theta^*) (\theta - \theta_0) \right) - 2\Gamma_{n0}(\theta) \\
&= P_n \left((\theta - \theta_0)' \nabla_1 \tau_n(z, \theta_0) \right) + \frac{1}{2} ((\theta - \theta_0)' [P_n(\nabla_2 \tau_n(z, \theta_0)) - 2V] (\theta - \theta_0)) + O(s_n |\theta - \theta_0|) + o(|\theta - \theta_0|^2) \\
&= \frac{1}{\sqrt{n}} (\theta - \theta_0)' W_n + \frac{1}{2} (\theta - \theta_0)' D_n (\theta - \theta_0) + O(s_n |\theta - \theta_0|) + o(|\theta - \theta_0|^2)
\end{aligned}$$

uniformly over $\alpha(1)$ neighborhood of θ_0 , where

$$\begin{aligned}
W_n &= \sqrt{n} P_n [(\mathcal{X} - \mathcal{X}_0) S(Y, \eta_0) g_0(\eta_0)], \\
D_n &= P_n \left[(\mathcal{X} - \mathcal{X}_0) (\mathcal{X} - \mathcal{X}_0)^T S(Y, \eta_0) g_0(\eta_0) \right] - 2\Delta
\end{aligned}$$

By Lemma 1, W_n converges in distribution to a $N(0, \Delta)$ random vector. By a weak law of large numbers, $D_n \rightarrow \Delta$ follows.

Lastly we show that

$$U_n r_n(\cdot, \cdot, \theta) = o_p(1), \quad (14)$$

Denote $H = \{I(y_1 > y_2)[r\{(1, \theta)(x_1 - x_2)/s\} - r\{(1, \theta_0)(x_1 - x_2)/s\}]: \theta \in \Theta, s \in [0, 1]\}$. Then H is Euclidean by Lemma 22 (ii) of [45]. By the dominated convergence theorem, $E(k_n(\cdot, \cdot, \theta)) \rightarrow 0$ as $\theta \rightarrow \theta_0$ and $s_n \rightarrow 0$. Thus $U_n k_n(\cdot, \cdot, \theta) = o_p(1)$ by following Theorem 3 of [26].

Together, (12), (13), (14) and $\frac{1}{n^2} \lambda_n \hat{\beta}_1^2 = o_p(1)$ imply (11). Thus completes the proof.

Appendix B: Derivation of the dual space criterion function and KKT conditions

In Section 3, we take the derivative of L_p with respect to β and ξ to get two sets of equations. Plug them back to L_p , we have

$$\begin{aligned}
L_p &= -\sum_p \left(h'_2(\eta_p^0) \mathbf{x}_p^T \boldsymbol{\beta} + \alpha_p \mathbf{x}_p^T \boldsymbol{\beta} - \frac{1}{2} \alpha_p \right) + \frac{1}{2} \boldsymbol{\beta}^T P_\lambda \boldsymbol{\beta} \\
&= -\sum_p (h'_2(\eta_p^0) + \alpha_p) \mathbf{x}_p^T P_\lambda^{-1} \sum_{p'} (h'_2(\eta_{p'}^0) + \alpha_{p'}) \mathbf{x}_{p'} + \sum_p \frac{1}{2} \alpha_p + \frac{1}{2} \sum_p (h'_2(\eta_p^0) + \alpha_p) \mathbf{x}_p^T P_\lambda^{-1} P_\lambda P_\lambda^{-1} \sum_{p'} (h'_2(\eta_{p'}^0) + \alpha_{p'}) \mathbf{x}_{p'} \\
&= -\frac{1}{2} \sum_{p,p'} (h'_2(\eta_p^0) + \alpha_p) (h'_2(\eta_{p'}^0) + \alpha_{p'}) \mathbf{x}_p^T P_\lambda^{-1} \mathbf{x}_{p'} + \frac{1}{2} \sum_p \alpha_p \\
&\propto -\frac{1}{2} \sum_{p,p'} (\alpha_p \alpha_{p'} + h'_2(\eta_p^0) \alpha_{p'} + \alpha_p h'_2(\eta_{p'}^0)) \mathbf{x}_p^T P_\lambda^{-1} \mathbf{x}_{p'} + \frac{1}{2} \sum_p \alpha_p \\
&= -\frac{1}{2} \sum_{p,p'} \alpha_p \alpha_{p'} \mathbf{x}_p^T P_\lambda^{-1} \mathbf{x}_{p'} - \frac{1}{2} \sum_{p,p'} 2h'_2(\eta_{p'}^0) \alpha_p \mathbf{x}_p^T P_\lambda^{-1} \mathbf{x}_{p'} + \frac{1}{2} \sum_p \alpha_p \\
&\propto -\boldsymbol{\alpha}^T Q \boldsymbol{\alpha} + \sum_p \left(1 - 2 \sum_{p'} h'_2(\eta_{p'}^0) \mathbf{x}_p^T P_\lambda^{-1} \mathbf{x}_{p'} \right) \alpha_p \\
&= -\boldsymbol{\alpha}^T Q \boldsymbol{\alpha} + \{ \mathbf{1} - 2Q \mathbf{h}'_2(\boldsymbol{\eta}^0) \}^T \boldsymbol{\alpha}.
\end{aligned}$$

Thus we have arrived at a dual space formulation of the optimization problem that is a function of $\boldsymbol{\alpha}$ only.

The KKT conditions associated with the above optimization programs are

$$\begin{aligned}
P_\lambda^{-1} \sum_p (h'_2(\eta_p^0) + \alpha_p) \mathbf{x}_p &= \boldsymbol{\beta} \\
\xi_p + \mathbf{x}_p^T \boldsymbol{\beta} - \frac{1}{2} &\geq 0 \\
\xi_p &\geq 0 \\
1 &\geq \alpha_p \geq 0 \\
\alpha_p \left(\xi_p + \mathbf{x}_p^T \boldsymbol{\beta} - \frac{1}{2} \right) &= 0 \\
(1 - \alpha_p) \xi_p &= 0
\end{aligned}$$

This set of conditions can be written more succinctly by noting that if $\alpha_p = 0$, then $\xi_p = 0$ and $\mathbf{x}_p^T \boldsymbol{\beta} - \frac{1}{2} \geq 0$; if $1 > \alpha_p > 0$, then $\xi_p = 0$ and $\mathbf{x}_p^T \boldsymbol{\beta} - \frac{1}{2} = 0$, or $\mathbf{x}_p^T \boldsymbol{\beta} = \frac{1}{2}$; if $\alpha_p = 1$, then $\mathbf{x}_p^T \boldsymbol{\beta} - \frac{1}{2} = -\xi_p \leq 0$. Combining them, we have the following expression of the KKT conditions

$$\begin{aligned}
\mathbf{x}_p^T \boldsymbol{\beta} - \frac{1}{2} &\geq 0 \quad \text{if } \alpha_p < 1 \\
\mathbf{x}_p^T \boldsymbol{\beta} - \frac{1}{2} &\leq 0 \quad \text{if } \alpha_p > 0
\end{aligned}$$

Figure 1.

(a) Different approximations of the empirical AUC loss function. (b) RAUC can be represented as a difference of two convex functions.

Figure 2. Sample training datasets from Section 4.1. Cases are plotted with filled circles, and controls are plotted with empty circles. Best linear combinations from three methods are shown as lines.

Test data AUC from Section 4.1, mean and se: Monte Carlo mean and standard error. SVM^l: SVM with linear kernel. RAUC^l: RAUC with linear kernel.

Table 1

Training data	Test data					
	No outliers			With outliers		
	logistic	SVM ^l	SAUC	RAUC ^l	logistic	SVM ^l SAUC RAUC ^l
No outliers						
mean	0.700	0.691	0.696	0.699	0.684	0.676 0.681 0.683
se	0.000	0.000	0.000	0.000	0.000	0.000 0.001 0.000
With outliers						
mean	0.639	0.648	0.667	0.681	0.632	0.639 0.657 0.669
se	0.001	0.001	0.002	0.001	0.001	0.001 0.001 0.001

Table 2

Test data AUC from Section 4.2, mean and se: Monte Carlo mean and standard error. SVM^f: SVM with RBF kernel. RAUC^f: RAUC with RBF kernel.

Training data	Test data							
	No outliers				With outliers			
	logistic	SAUC	SVM ^f	RAUC ^f	logistic	SAUC	SVM ^f	RAUC ^f
No outliers								
mean	0.675	0.674	0.828	0.840	0.685	0.683	0.789	0.796
se	0.000	0.000	0.001	0.000	0.000	0.000	0.001	0.000
With outliers								
mean	0.675	0.674	0.790	0.817	0.685	0.684	0.795	0.816
se	0.000	0.000	0.001	0.001	0.000	0.000	0.001	0.000

Table 3

Test data AUC for the RV144 data example, mean and se: mean and standard error from repeated random sub-sampling validation.

	logistic	SAUC	SVM ^f	RAUC ^f	SVM ^r	RAUC ^r
mean	0.627	0.609	0.624	0.621	0.588	0.656
se	0.003	0.002	0.003	0.002	0.002	0.003