Statistica Sinica Preprint No: SS-2024-0195	
Title	Maximizing Area Under the Receiver Operating Characteristic Curve for Biomarker Combination
Manuscript ID	SS-2024-0195
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0195
Complete List of Authors	Yuxuan Chen and Yijian Huang
Corresponding Authors	Yuxuan Chen
E-mails	yuxuan.chen@emory.edu

Maximizing Area Under the Receiver Operating Characteristic Curve for Biomarker Combination

Yuxuan Chen and Yijian Huang

Department of Biostatistics and Bioinformatics, Emory University

Abstract:

Multiple biomarkers are often combined for more accurate disease diagnosis. For this purpose, one popular performance metric is the area under the receiving operating characteristic (ROC) curve (AUC). Optimizing the empirical AUC over linear combinations of biomarkers, however, faces two primary challenges. First, AUC is scale-invariant to the linear combinations, creating difficulties in both the computation and asymptotic study. Most available approaches actually consider a restricted problem by setting one coefficient to a constant. Second, the empirical AUC is piecewise-constant and standard gradient-based computational algorithms are not applicable. Existing methods maximize kernel-smoothed AUC instead, but they can be sensitive to bandwidth choice. In this article, we tackle these challenges by developing a new empirical AUC maximization method. Computationally efficient algorithms are provided for both the point and variance estimation of the estimated combination coefficients. Simulation studies show good computational and statistical performance of the proposed methods. An illustration is provided with a clinical application.

Key words and phrases: classification, kernel smoothing, optimization algorithm, Receiver Operating Characteristic curve, variance estimation

1. Introduction

With advances in high-throughput sequencing and molecular profiling technologies, many biomarkers have been identified for the purpose of disease diagnosis. As a single biomarker may not be sufficiently informative, combining multiple biomarkers holds the promise for improved accuracy. For this purpose, likelihood-based methods such as logistic regression are widely adopted. Despite their good computational properties and wide acceptance, such a combination may be suboptimal in the case of model misspecification. As an alternative and more robust method, Pepe et al. (2006) considered the area under the receiver operating characteristic curve (AUC) and suggested combining biomarkers to maximize the empirical AUC. AUC is a popular performance metric and can be interpreted as the probability that a diseased individual has a larger biomarker combination than a healthy individual.

Unfortunately, the empirical AUC maximization poses two major challenges. The first one pertains to the scale-invariance property of AUC to the combination coefficient. To address this identifiability issue, the majority of

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0195

existing methods designate one biomarker as an anchor to fix its coefficient to a non-zero constant (Pepe and Thompson, 2000; Vexler et al., 2006; Ma and Huang, 2007; Zhang et al., 2018; Chen et al., 2015). However, this is a restricted problem since it requires a priori knowledge of a biomarker to have a non-zero coefficient with a certain sign. A better solution, without compromising the generality, is to impose a norm constraint on the combination coefficient. Unfortunately, such a norm constraint is difficult to deal with computationally; see ad hoc solutions in Lin et al. (2011) and Fong et al. (2016). Furthermore, the variance estimation becomes difficult with the degenerate distribution of the estimated coefficients.

Another challenge stems from the fact that the empirical AUC is piecewise constant. Standard optimization algorithms, such as those gradient-based, are not applicable. Most existing methods maximize a kernel-smoothed empirical AUC instead by replacing the indicator function in the empirical AUC with a smoothed kernel function, e.g., Gaussian kernel by Vexler et al. (2006) and Lin et al. (2011), sigmoid kernel by Ma and Huang (2007), and ramp kernel by Fong et al. (2016). Nevertheless, these estimators can be sensitive to the choice of bandwidth. A small bandwidth may not sufficiently improve the computational properties, but a large one can result in a statistically different estimator. Sound procedures for bandwidth choice are typically lacking. Zhang et al. (2018) proposed a different smoothing technique that can be applied to the empirical AUC maximization. Their self-induced smoothing introduces random perturbation to the combination coefficient, with the amount of perturbation adapted to the data to result in asymptotically equivalent estimation. However, the computation can be burdensome, and convergence is not always guaranteed for small sample size. An alternative approach is to employ optimization techniques like simulated annealing to directly maximize the empirical AUC. Simulated annealing does not require the objective function to be differentiable or continuous, making it suitable for optimizing piecewise constant functions. However, it may suffer from slow convergence and high computational cost, and its performance is sensitive to parameter choices like the cooling schedule, which can be challenging to tune.

For inference of the combination coefficients, sandwich variance estimation for coefficients as considered in Ma and Huang (2007) and Fong et al. (2016) may also be sensitive to the bandwidth choice and may not perform well, particularly with small to moderate sample sizes. Bootstrap is applicable, but the computation is intensive (Ma and Huang, 2007). The self-induced smoothing of Zhang et al. (2018) provides variance estimation simultaneously with point estimation. However, again, the iterative procedure can be computationally intensive.

In this article, we develop a novel method for the empirical AUC maximization problem to tackle the various issues associated with the existing methods. We work with the general problem, where a norm constraint is imposed on the combination coefficients. As an innovation, we introduce an equivalent unconstrained reformulation of the constrained optimization problem to facilitate both the point estimation computation and variance estimation. With point estimation, a novel algorithm is designed for efficient and robust computation by developing a sequence of smoothed objective functions that converge to the target one. Furthermore, we present a novel sandwich-type variance estimate with efficient computation.

The rest of the article is organized as follows. Section 2 describes the proposed estimation procedure for the combination coefficients. Section 3 shows the numerical performance of our proposal in comparison with existing methods through simulations and a real data example. Section 4 concludes with a discussion. Assumptions and technical proofs are deferred to the Appendix.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0195

2. The Proposed Point and Variance Estimation Methods

Write $Y = 1, 0$ as the presence or absence of a disease, respectively, and $\mathbf{X} \in \mathbb{R}^d$ as a vector of d biomarkers of interest for $d \geq 2$. In a cohort study, the observed data consist of n independent and identically distributed (iid) replicates of $(Y, \mathbf{X})$: $(Y_i, \mathbf{X}_i), i = 1, \ldots, n$. Consider a linear combination $\boldsymbol{b}^T\mathbf{X}$ with coefficient $\boldsymbol{b} \in \mathbb{R}^d$. Adopt the convention that a larger biomarker combination is associated with positive disease diagnosis. With combination coefficient $\boldsymbol{b}$, the AUC is equal to $A(\boldsymbol{b}) = Pr(\boldsymbol{b}^T\mathbf{X}_1 > \boldsymbol{b}^T\mathbf{X}_2 | Y_1 = 1, Y_2 = 0)$. The empirical AUC is given by:

$$\widehat{A}(\boldsymbol{b}) = \frac{1}{N_1 N_0} \sum_{i \neq j} I(Y_i > Y_j) I(\boldsymbol{b}^T\mathbf{X}_i > \boldsymbol{b}^T\mathbf{X}_j),$$

where N_1 and N_0 are the numbers of the diseased and healthy individuals, respectively.

Since both $A(\boldsymbol{b})$ and $\widehat{A}(\boldsymbol{b})$ are scale-invariant to $\boldsymbol{b}$, we impose $\|\boldsymbol{b}\|_2 = 1$ for identifiability without loss of generality. The optimal combination is thus given by $\boldsymbol{\beta}_0 = \text{argmax}_{\|\boldsymbol{b}\|_2=1} A(\boldsymbol{b})$. For its estimation, we consider the empirical AUC maximizer,

$$\widehat{\boldsymbol{\beta}} = \underset{\|\boldsymbol{b}\|_2=1}{\text{argmax}}\, \widehat{A}(\boldsymbol{b}). \tag{2.1}$$

Han (1987) and Sherman (1993) studied this estimator or a related one under a semiparametric model. Their strong consistency and asymptotic normality results of the estimator can be extended to the current nonparametric set-up.

2.1 Point Estimation through Diminishing Smoothing

The norm constraint $\|\boldsymbol{b}\|_2 = 1$ causes computational challenges, which have not been well addressed in existing methods, e.g., Fong et al. (2016). By exploiting the scale-invariance of $\widehat{A}(\boldsymbol{b})$ to $\boldsymbol{b}$, we obtain a novel unconstrained reformulation through penalization:

$$\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{b}} \widehat{A}(\boldsymbol{b}) - w(\|\boldsymbol{b}\|_2 - 1)^2. \tag{2.2}$$

where w is a positive constant. This new formulation also facilitates the asymptotic study on weak convergence as the Hessian of the objective function is no longer singular in limit.

Proposition 1. *Denote*

$$\begin{aligned}\tau_0(y,\boldsymbol{x};\boldsymbol{b}) &= \mathrm{E}\left[I(y>Y)I\{(\boldsymbol{x}-\mathbf{X})^T\boldsymbol{b}>0\}\right.\\ &\qquad\left.+\,I(y<Y)I\{(\boldsymbol{x}-\mathbf{X})^T\boldsymbol{b}<0\}\right],\\ \mathbf{H}_0 &= 2\{\rho(1-\rho)\}^{-1}\mathrm{E}\left\{\nabla_2\tau_0(Y,\mathbf{X};\boldsymbol{\beta}_0)\right\},\\ \textit{and}\quad \mathbf{V}_0 &= \{\rho(1-\rho)\}^{-2}\mathrm{E}\left\{\nabla\tau_0(Y,\mathbf{X};\boldsymbol{\beta}_0)\right\}^{\otimes 2},\end{aligned}$$

where $\rho = \mathrm{E}(Y)$ with $\rho \in (0,1)$. If Assumptions 1 to 3 in the Appendix hold, then

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\xrightarrow{d} N(\mathbf{0},\boldsymbol{\Sigma}_0),$$

where $\boldsymbol{\Sigma}_0 = \mathbf{H}_P^{-1}\mathbf{V}_0\mathbf{H}_P^{-1}$ with $\mathbf{H}_P = \mathbf{H}_0 - 2w\boldsymbol{\beta}_0^{\otimes 2}$.

Another feature of the problem that contributes to the computational difficulty is the indicator function involved in $\widehat{A}(\boldsymbol{b})$. To address that, we develop a sequence of smoothed objective functions that converge to that in (2.2). Such a smoothed objective function is obtained through kernel smoothing, similar to existing methods, e.g., Ma and Huang (2007). However, our approach is distinct in adopting the sequence to overcome the need for bandwidth choice. Specifically, for a given smoothing parameter $\sigma > 0$, we approximate the indicator function $I(x > 0)$ by $g_\sigma(x) - f_\sigma(x)$, where

$g_\sigma(x)$ and $f_\sigma(x)$ are both continuous, piecewise-quadratic, and convex; see Figure 1:

$$g_\sigma(x) = \begin{cases} -\frac{1}{\sigma}x - \frac{1}{4} & \text{if } x \in (-\infty, -\sigma], \\ \frac{1}{2\sigma^2}x^2 + \frac{1}{4} & \text{if } x \in (-\sigma, 0], \\ \frac{1}{\sigma}x + \frac{1}{4} & \text{if } x \in (0, \infty), \end{cases}$$

and

$$f_\sigma(x) = \begin{cases} -\frac{1}{\sigma}x - \frac{1}{4} & \text{if } x \in (-\infty, 0], \\ \frac{1}{2\sigma^2}x^2 - \frac{1}{4} & \text{if } x \in (0, \sigma], \\ \frac{1}{\sigma}x - \frac{3}{4} & \text{if } x \in (\sigma, \infty). \end{cases}$$

This approximation function is specifically designed to facilitate the subsequent algorithm development. Then, $\widehat{A}(\boldsymbol{b})$ is approximated by $\widetilde{A}_\sigma(\boldsymbol{b})$:

$$\widetilde{A}_\sigma(\boldsymbol{b}) = \frac{1}{N_1 N_0} \sum_{i \neq j} I(Y_i > Y_j)\{g_\sigma(\boldsymbol{b}^T\mathbf{X}_i - \boldsymbol{b}^T\mathbf{X}_j) - f_\sigma(\boldsymbol{b}^T\mathbf{X}_i - \boldsymbol{b}^T\mathbf{X}_j)\}.$$

Consequently, we have

$$\widetilde{\boldsymbol{\xi}}_\sigma = \arg\max_{\boldsymbol{b}} \widetilde{A}_\sigma(\boldsymbol{b}) - w(\|\boldsymbol{b}\|_2 - 1)^2. \tag{2.3}$$

Since $\widetilde{A}_\sigma(\boldsymbol{b})$ is no longer scale-invariant to $\boldsymbol{b}$, $\|\widetilde{\boldsymbol{\xi}}_\sigma\|_2$ is not necessarily 1.

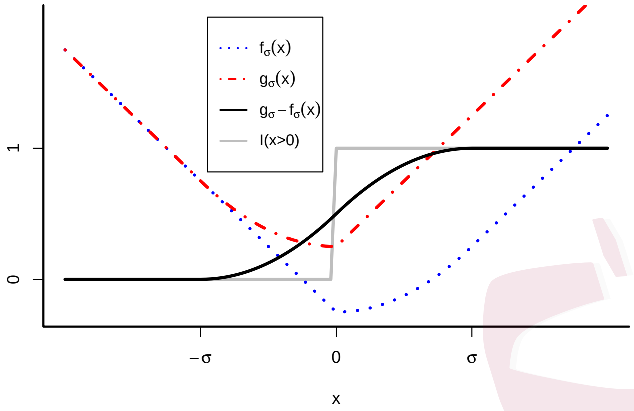

Figure 1: Approximation of the indicator function $I(x > 0)$ by the difference of two continuous, piecewise-quadratic, and convex functions $g_\sigma(x)$ and $f_\sigma(x)$

Nevertheless, with $\widetilde{A}_\sigma(\boldsymbol{b}) \in [0, 1]$, we have $1 - w(\|\widetilde{\boldsymbol{\xi}}_\sigma\|_2 - 1)^2 \geq 0$, that is, $|\|\widetilde{\boldsymbol{\xi}}_\sigma\|_2 - 1| \leq 1/\sqrt{w}$. Provided $w > 1$, we have

$$\widetilde{\boldsymbol{\xi}}_\sigma = \arg\max_{\|\boldsymbol{b}\|_2 = s} \widetilde{A}_\sigma(\boldsymbol{b})$$

for some data-dependent $s \in [1 - 1/\sqrt{w}, 1 + 1/\sqrt{w}]$. Equivalently, $\widetilde{\boldsymbol{\beta}}_\sigma$, defined as $\widetilde{\boldsymbol{\xi}}_\sigma / \|\widetilde{\boldsymbol{\xi}}_\sigma\|_2$, has an alternative representation

$$\widetilde{\boldsymbol{\beta}}_\sigma = \arg\max_{\|\boldsymbol{b}\|_2 = 1} \widetilde{A}_{s^{-1}\sigma}(\boldsymbol{b}).$$

A similar technique was implemented in Huang and Sanda (2022) to address

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0195

scale-invariance in their optimization problem.

Theorem 1. *Suppose that $w > 1$. If Assumptions 1–3 in the Appendix hold and $\sigma = o(n^{-1/2})$, then $\widetilde{\boldsymbol{\beta}}_\sigma$ is asymptotically equivalent to $\widehat{\boldsymbol{\beta}}$ in the sense that $\widetilde{\boldsymbol{\beta}}_\sigma = \widehat{\boldsymbol{\beta}} + o_p(n^{-1/2})$.*

With a sufficiently small σ, Theorem 1 shows that the difference between $\widetilde{\boldsymbol{\beta}}_\sigma$ and $\widehat{\boldsymbol{\beta}}$ is asymptotically negligible. We shall let σ approach 0 in our computation algorithm so as to spare the need to choose a small value for σ.

Now, focus on the optimization algorithm for problem (2.3). We consider an equivalent minimization problem as being more standard in the optimization literature:

$$\min_{\boldsymbol{b}} -\widetilde{A}_\sigma(\boldsymbol{b}) + w(\|\boldsymbol{b}\|_2 - 1)^2. \tag{2.4}$$

Notice that the penalty can be written as the difference of $w(\|\boldsymbol{b}\|_2^2 + 1)$ and $2w\|\boldsymbol{b}\|_2$, both of which are convex. On the other hand, $\widetilde{A}_\sigma(\boldsymbol{b})$ is also a difference of two convex functions by design. Therefore, the objective function in problem (2.4) is readily written as the difference of two convex functions $F_\sigma(\boldsymbol{b})$ and $G_\sigma(\boldsymbol{b})$:

$$\begin{aligned} F_\sigma(\boldsymbol{b}) &= \frac{1}{N_1 N_0}\sum_{i=1}^{N_1}\sum_{j=1}^{N_0} f_\sigma(\boldsymbol{b}^T\mathbf{X}_i - \boldsymbol{b}^T\mathbf{X}_j) + w(\boldsymbol{b}^T\boldsymbol{b}+1). \\ G_\sigma(\boldsymbol{b}) &= \frac{1}{N_1 N_0}\sum_{i=1}^{N_1}\sum_{j=1}^{N_0} g_\sigma(\boldsymbol{b}^T\mathbf{X}_i - \boldsymbol{b}^T\mathbf{X}_j) + 2w\|\boldsymbol{b}\|_2, \end{aligned}$$

Consequently, we can implement the concave-convex procedure (CCCP) by Yuille and Rangarajan (2003), which constitutes the core of Algorithm 1. At each iteration, we convexify the objective function in (2.4) by linearizing G_σ, i.e., replacing it by its tangent plane at the current coefficient estimate. The resulting unconstrained convex problem can be solved with the Newton method or quasi-Newton method. The current combination coefficient estimate is updated with the optimizer with guaranteed improvement of the objective function. Such iterations are repeated until the objective function cannot be further improved.

We have so far focused on the optimization algorithm for problem (2.4) with a fixed σ. To approach the original objective function, a sequence of decreasing σ values is adopted in the outer loop of Algorithm 1. With a larger σ, the smoothed objective function may have fewer local optima. As σ value is gradually reduced, the estimator might more likely approach the global optimizer.

Our kernel function as given in (2.4) facilitates the proposed computation in several ways. First, its representation as the difference of two simple convex functions lends itself for the implementation of the CCCP algorithm. Second, the kernel is differentiable so that efficient gradient-based optimization algorithms can be applied to the convexified objective function. Third, the kernel function has a finite support which can be exploited to reduce the computational burden of calculating the convexified objective function and its gradient. Because many pairs of biomarker combinations may not contribute to the change in the convexified objective function, a considerable number of unnecessary comparisons between combinations can be avoided by sorting the biomarker combinations of all subjects.

2.2 Sandwich-type Variance Estimation

Variance estimation for $\widehat{\boldsymbol{\beta}}$ is challenging. First, even $A(\boldsymbol{b})$ has a singular Hessian because of its scale invariance to $\boldsymbol{b}$. Standard M-estimation theory thus cannot be applied directly to the maximizer of $\widehat{A}(\boldsymbol{b})$. Our incorporation of the penalty function in problem (2.2) offers a solution to this problem, since the limit of the objective function now has a Hessian of full rank.

Another challenge with variance estimation stems from the nonsmoothness of $\widehat{A}(\boldsymbol{b})$. As such, $\mathbf{H}_0$ cannot be estimated through direct differentiation

Algorithm 1 Pseudo code for problem (2.2)

```
Given an initial combination coefficient $\boldsymbol{b}_0$, initial value for $\sigma$, and shrinkage factor $\alpha$.
Set $k \leftarrow 0$.
repeat
    repeat                                  ▷ Concave-convex procedure with fixed $\sigma$
        1. Convexify.
           Form
                $$G^*_\sigma(\boldsymbol{b}; \boldsymbol{b}_k) = G_\sigma(\boldsymbol{b}_k) + \nabla G_\sigma(\boldsymbol{b}_k)^T(\boldsymbol{b} - \boldsymbol{b}_k).$$
        2. Solve. Set $\boldsymbol{b}_{k+1}$ to the solution of
                $$\min_{\boldsymbol{b}} F_\sigma(\boldsymbol{b}) - G^*_\sigma(\boldsymbol{b}; \boldsymbol{b}_k).$$
        3. Update inner loop. $k \leftarrow k + 1$.
    until Sufficiently small decrease in objective function (2.4) with current $\sigma$
                                            ▷ End of inner loop
    Shrink $\sigma$ by the factor $\alpha$.
until Sufficiently small decrease in objective function (2.2)
```

of $\widehat{A}(\boldsymbol{b})$. To address that, we adapt the self-induced smoothing approach of Zhang et al. (2018), which incorporates data-induced smoothing at the right level to allow sandwich-type variance estimation. Nevertheless, our adaptation has two distinctive features. As our formulation of the problem is more general, our variance matrix of $\widehat{\boldsymbol{\beta}}$ is nearly singular, which needs to be accommodated. On the other hand, unlike Zhang et al. (2018), the point estimate $\widehat{\boldsymbol{\beta}}$ is fixed in our procedure to have computational advantages.

Given a positive definite matrix $\boldsymbol{\Sigma}$, denote $\sigma_{ij} = \sqrt{\mathbf{X}_{ij}^T \boldsymbol{\Sigma} \mathbf{X}_{ij}}$, where $\mathbf{X}_{ij} = \mathbf{X}_i - \mathbf{X}_j$. Let Φ be the standard normal distribution function. The objective function with self-induced smoothing can be written as:

$$L(\widehat{\boldsymbol{\beta}}, \boldsymbol{\Sigma}) = \frac{1}{N_1 N_0} \sum_{i=1}^{N_1} \sum_{j=1}^{N_0} \Phi\left(\frac{\sqrt{n} \mathbf{X}_{ij}^T \widehat{\boldsymbol{\beta}}}{\sigma_{ij}} \right) - w(\|\widehat{\boldsymbol{\beta}}\|_2 - 1)^2.$$

Then $\boldsymbol{\Sigma}_0$ can be estimated by the sandwich-type variance estimation:

$$\widehat{\boldsymbol{\Sigma}} = \widehat{\mathbf{H}}_P^{-1}(\widehat{\boldsymbol{\beta}}, \boldsymbol{\Sigma}) \times \widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}}, \boldsymbol{\Sigma}) \times \widehat{\mathbf{H}}_P^{-1}(\widehat{\boldsymbol{\beta}}, \boldsymbol{\Sigma}), \tag{2.5}$$

where

$$\widehat{\mathbf{H}}_P(\widehat{\boldsymbol{\beta}}, \boldsymbol{\Sigma}) = \nabla_2 L(\widehat{\boldsymbol{\beta}}, \boldsymbol{\Sigma}),$$

$$\widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}}, \boldsymbol{\Sigma}) = \frac{n}{N_1^2 N_0^2} \sum_{i=1}^{n} \left[\sum_{j=1}^{n} \left\{ sgn(Y_i - Y_j) \times \phi\left(\frac{\sqrt{n}\mathbf{X}_{ij}^T \widehat{\boldsymbol{\beta}}}{\sigma_{ij}} \right) \frac{\sqrt{n}\mathbf{X}_{ij}}{\sigma_{ij}} \right\} \right]^{\otimes 2}.$$

Theorem 2. *Let $\widehat{\boldsymbol{\beta}}$ be the maximizer of empirical AUC, and $\widehat{\boldsymbol{\Sigma}}$ as defined in Equation (2.5). Suppose Assumptions 1 to 3 in the Appendix hold. Then, for any fixed positive definite matrix $\boldsymbol{\Sigma}$, $\widehat{\boldsymbol{\Sigma}}$ converges in probability to $\boldsymbol{\Sigma}_0$, the limiting variance-covariance matrix of $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$.*

Because of the norm constraint, $\boldsymbol{\Sigma}_0$ is of rank $d-1$ and its column vectors do not span $\boldsymbol{\beta}_0$. Thus, $\widehat{\boldsymbol{\Sigma}}$ is nearly singular since it is consistent for $\boldsymbol{\Sigma}_0$. Therefore, we consider to use $\widehat{\boldsymbol{\Sigma}} + \widehat{\boldsymbol{\beta}}^{\otimes 2}$ as $\boldsymbol{\Sigma}$ above for the purpose of smoothing. By adding $\widehat{\boldsymbol{\beta}}^{\otimes 2}$, $\boldsymbol{\Sigma}$ has full rank. With this choice, we develop an iterative procedure illustrated in Algorithm 2. The estimation procedure begins with a working covariance matrix, such as the estimated variance matrix of the estimated coefficients from logistic regression, and iteratively updates the variance estimate using (2.5).

Algorithm 2 Sandwich-type Algorithm for Variance Estimation

```
Given a coefficient estimate $\widehat{\boldsymbol{\beta}}$, an initial covariance matrix $\widehat{\boldsymbol{\Sigma}}^{(0)}$; set $k \leftarrow 0$.
repeat
    1. Compute $\widehat{\mathbf{H}}_P(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}}^{(k)} + \widehat{\boldsymbol{\beta}}^{\otimes 2})$ and $\widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}}^{(k)} + \widehat{\boldsymbol{\beta}}^{\otimes 2})$.
    2. Update $\widehat{\boldsymbol{\Sigma}}^{(k)}$ using Equation (2.5).
    3. Update iteration. $k \leftarrow k + 1$.
until $\widehat{\boldsymbol{\Sigma}}^{(k+1)}$ and $\widehat{\boldsymbol{\Sigma}}^{(k)}$ are sufficiently close to each other.
```

3. Numerical Studies

We implemented Algorithm 1 by adopting the Broyden–Fletcher–Goldfarb–Shanno (BFGS) quasi-Newton method for convex optimization (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970) from R `optim` function. The maximum likelihood estimator from logistic regression is used as the initial coefficient, the weight w is set to 2, and the shrinking factor α is set to 0.8. The code for the proposed estimator is available as an R package on the first author's website `https://github.com/yuxuanchn/maxAUC`.

Six biomarker combination estimation methods were included in the numerical studies for comparison of point estimation. Logistic regression (LR) is a standard and widely adopted method. We also considered four representative empirical AUC maximization methods: the Gaussian-smoothed AUC (GAUC), implemented with bandwidth choices suggested by Vexler et al. (2006) and R code from their subsequent work (Chen et al., 2015); the sigmoid-smoothed AUC (SAUC), with bandwidth choice proposed by Ma and Huang (2007); the ramp AUC (RAUC) (Fong et al., 2016) as implemented in R package `aucm`; and the smoothed maximum rank correlation estimator (SMRCE) proposed by Zhang et al. (2018). We did not have access to the original code for Ma and Huang (2007) and SMRCE and wrote the code using the BFGS quasi-Newton method from R `optim` function. As these methods require the designation of an anchor, we selected the biomarker with the largest absolute coefficient from logistic regression. Additionally, we included empirical AUC maximization using simulated annealing (SANN) as implemented in the R function `optim` (Bélisle, 1992).

For variance estimation, we compared our proposal with two existing methods. One is SMRCE (Zhang et al., 2018), which simultaneously produces the variance and point estimation. The other is the weighted bootstrap method adopted by Ma and Huang (2007) for their SAUC estimator, where the random weights were generated from $Beta(0.125, 1.125)$.

3.1 Simulation

In the simulation studies, the disease status marginally followed a Bernoulli distribution with a prevalence of 30% and three biomarkers were considered for combination. Given the disease status, biomarkers followed a conditional distribution that was considered in the simulation studies of Huang and Sanda (2022), as motivated from cancer studies. For healthy individuals, the biomarkers were independent and identically distributed, following a standard normal distribution. On the other hand, two distributions of the biomarkers of diseased individuals were considered:

> *Scenario I.* The biomarkers of diseased individuals were independent and normally distributed with mean 0.9 and variance 1.
>
> *Scenario II.* The biomarkers of diseased individuals were independent and normally distributed, following a mixture of independent normal distributions. With probability 2/3, the means are (1.7, 1,7, 0) and variances (0.5, 2, 1); with probability 1/3, the means are (0, 0, 1.7) with variances of 1.

In Scenario I, all three biomarkers of diseased individuals tend to be larger than those of healthy individuals. Scenario II mimics two disease subtypes where the elevation of the first two biomarkers is associated with one subtype, and that of the third biomarker is associated with the other. The logistic regression model holds under Scenario I, but not under Scenario II. Sample sizes of 100, 200, 500, 2000 were considered. For each set-up, results for 1000 random samples were obtained. The simulations were conducted

Table 1: Simulation study results: Computational performance of coefficient estimation

n	100		200		500		2000	
Method	A	T	A	T	A	T	A	T
					Scenario I			
PROP	87.64	0.07	87.07	0.12	86.73	0.42	86.52	4.95
SAUC	87.47	0.92	87.02	1.83	86.72	2.75	86.52	10.83
GAUC	87.56	0.42	87.01	0.74	86.71	2.03	86.52	12.10
RAUC	87.35	0.28	86.95	3.04	-	-	-	-
SMRCE	87.34	0.03	86.94	0.11	86.70	0.63	86.52	9.41
SANN	87.33	0.09	86.86	0.17	86.61	0.73	86.48	8.14
					Scenario II			
PROP	87.77	0.07	86.75	0.12	86.40	0.41	86.22	5.14
SAUC	87.59	0.88	86.69	1.57	86.38	4.20	86.22	10.89
GAUC	87.68	0.64	86.69	1.10	86.37	2.49	86.22	21.07
RAUC	87.48	0.32	86.62	3.49	-	-	-	-
SMRCE	87.43	0.03	86.58	0.12	86.35	0.64	86.22	9.55
SANN	87.08	0.08	86.62	0.17	86.29	0.71	86.11	9.80

A: Optimized empirical AUC (×100); T: computation time (seconds).
PROP: proposed method; SAUC: Sigmoid-smoothed AUC; GAUC: Gaussian-smoothed AUC; RAUC: ramp AUC; SMRCE: smoothed maximum rank correlation estimator; SANN: simulated annealing. RAUC was not performed for larger sample sizes due to lengthy running time.

on a 2023 Mac mini with an M2 chip and 16 GB memory.

We start with computational performance, with the resulting maximal empirical AUC and computation time shown in Table 1. As all methods intend to maximize the empirical AUC, one achieving a larger maximal empirical AUC is regarded better. In all settings, our algorithm attained the largest empirical AUC on average. Meanwhile, our method generally had shorter computation time. The advantage became more substantial as the number of biomarkers and sample size increased.

Next, we consider the statistical performance of coefficient estimation. Table 2 shows the bias and standard deviation of the estimated coefficients. Under Scenario I, the logistic regression model holds and the associated coefficient estimate is asymptotically efficient. In this case, our proposed coefficient estimate had slightly inflated standard deviation, but was comparable to other AUC maximization methods. In Scenario II, the bias of all methods except logistic regression decreased and approached 0 as the sample size increased. This is not surprising since the logistic regression model no longer holds. The simulations also suggested that the logistic regression coefficients do not converge to the optimal combination coefficients. Across all set-ups, simulated annealing had larger bias and standard deviation, which may not be surprising as it is a general-purpose algorithm.

Finally, we compare our proposed variance estimation procedure with the weighted bootstrap in Ma and Huang (2007) and SMRCE. As both comparing methods designate an anchor biomarker, their variance estimations are transformed by the delta method for comparison with the proposed method. Table 3 shows the standard error and coverage probability of 95% confidence interval for each coefficient, along with the computation time. As expected, the standard error reduced and the coverage rate approached the nominal level with the increase of sample size. The standard error and

Table 2: Simulation study results on coefficient estimation

	Scenario I						Scenario II					
	β_1		β_2		β_3		β_1		β_2		β_3	
Method	B	SD	B	SD	B	SD	B	SD	B	SD	B	SD
							n=100					
PROP	-25	150	-25	159	-11	147	-39	154	11	149	-19	137
LR	-21	142	-25	151	-8	135	-37	144	42	140	-43	139
SAUC	-25	151	-29	160	-7	145	-38	155	18	147	-24	134
GAUC	-26	155	-25	159	-12	147	-40	155	11	151	-19	137
RAUC	-26	153	-20	155	-14	145	-23	148	1	150	-32	145
SMRCE	-11	136	-26	142	-12	128	-35	142	15	127	-18	135
SANN	-42	204	-37	202	-32	204	-49	207	24	187	-64	191
							n=200					
PROP	-10	105	-10	107	-10	110	-7	112	-12	121	-16	100
LR	-5	102	-10	101	-12	103	-10	108	37	109	-50	104
SAUC	-9	106	-10	107	-11	111	-7	112	-8	120	-18	102
GAUC	-9	107	-10	110	-12	112	-6	114	-14	121	-15	101
RAUC	-9	106	-11	112	-12	110	-11	112	0	114	-18	100
SMRCE	-5	96	-6	95	-13	101	-5	101	-4	107	-18	99
SANN	-20	136	-18	137	-10	134	-28	159	36	144	-57	149
							n=500					
PROP	-4	75	-5	70	-4	67	-13	70	5	69	-0	64
LR	-4	69	-5	63	-2	64	-16	63	53	62	-34	66
SAUC	-4	74	-5	70	-4	66	-12	70	5	69	-0	62
GAUC	-4	74	-5	69	-4	66	-12	69	4	69	-0	62
SMRCE	-4	68	-4	62	-3	61	-12	60	7	60	-1	61
SANN	-10	98	-8	98	-7	94	-24	138	39	107	-42	118
							n=2000					
PROP	0	34	-1	35	-2	35	-1	33	0	36	-1	31
LR	2	31	-1	33	-3	33	-7	30	49	33	-34	32
SAUC	1	33	-1	35	-2	35	-1	33	0	35	-1	31
GAUC	1	34	-1	34	-2	35	-1	33	0	35	-1	31
SMRCE	1	31	-1	32	-3	33	-1	31	0	33	-1	30
SANN	-2	48	-4	49	-1	50	-15	112	41	79	-39	91

B: bias (×1000); SD: standard deviation (×1000).
PROP: proposed method; LR: logistic regression; SAUC: Sigmoid-smoothed AUC; GAUC: Gaussian-smoothed AUC; RAUC: ramp AUC; SMRCE: smoothed maximum rank correlation estimator; SANN: simulated annealing. RAUC was not performed for larger sample sizes due to lengthy running time.

coverage probability were comparable across all three methods. However, our proposed method took much shorter computation time, and this computational advantage became more prominent as the sample size increased.

These reported simulations are only part of the studies that we performed. Additional simulation results featuring more biomarkers are included in the Supplementary Material. The computational and statistical performance of our proposed method relative to the existing ones remains similar.

3.2 Real Data

We applied the proposed method to the Pima Indians Diabetes Study, which examined the diagnosis of diabetes using eight variables in Pima indians aged 21 years and older (Smith et al., 1988). The same dataset was analyzed by Ma and Huang (2007). The data consisted of 268 diabetic patients and 500 nondiabetic individuals. These variables under consideration are easily measurable and thus can be used in emergency situations and patient self-care. We first compared the estimated optimal empirical AUC of each method. Our proposed method achieved the maximal empirical AUC (0.8410). The other empirical AUC maximization methods attained similar empirical AUC (GAUC: 0.8405; SAUC: 0.8405; SMRCE: 08397), all of

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0195

Table 3: Simulation study result on variance estimation

	Scenario I							Scenario II						
	β_1		β_2		β_3			β_1		β_2		β_3		
Method	E	C	E	C	E	C	T	E	C	E	C	E	C	T
							n=100							
PROP	150	89.6	150	89.6	150	90.8	0.01	152	86.4	148	92.0	134	89.2	0.01
SMRCE	133	88.8	129	88.0	132	91.6	0.03	124	83.2	115	88.8	133	92.4	0.03
WB	148	91.6	150	89.8	150	91.8	2304	141	90.0	148	91.0	139	90.6	2750
							n=200							
PROP	103	92.4	103	91.2	104	92.4	0.02	103	88.8	105	87.6	98	90.0	0.02
SMRCE	95	93.6	97	92.8	96	93.6	0.11	92	90.4	88	87.6	98	92.4	0.12
WB	108	91.8	108	92.6	108	91.8	5825	106	91.2	111	92.2	98	94.0	7443
							n=500							
PROP	66	91.2	66	92.8	66	92.0	0.09	66	91.6	67	94.0	60	89.6	0.09
SMRCE	64	94.4	64	95.2	63	95.2	0.63	63	94.8	60	92.4	62	93.2	0.64
							n=2000							
PROP	33	95.2	33	93.6	33	93.2	1.37	33	95.2	34	94.8	30	92.8	1.42
SMRCE	33	96.8	33	96.0	33	94.8	9.41	33	95.2	32	94.4	32	96.0	9.55

E: Standard error (×1000); C: coverage rate (%); T: average computation time (seconds).

PROP: proposed sandwich-type variance estimation method; SMRCE: smoothed maximum rank correlation estimation method; WB: Weighted Bootstrap method. Weighted Bootstrap method was not performed for larger sample sizes due to lengthy running time.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0195

which were slightly higher than the empirical AUC from LR (0.8394). The RAUC, however, did not converge and attained a smaller estimate (0.7340). The simulated annealing achieved the same emprical AUC as LR (0.8394). The slight improvement with our proposal might be due to the relatively large sample size, which is consistent with our simulation results.

The coefficient estimation and inference results were summarized in Table 4. The proposed method identified five variables as significant, where *Diabetes pedigree function* has the smallest p-value. For comparison, the logistic regression model and MH yield similar coefficient estimation. However, the inference results were different. For example, *Diabetes pedigree function* has the smallest p-value in both weighted bootstrap and the proposed method, but *Glucose concentration* has the smallest p-value in the logistic model. This difference in estimation results from the difference in their model assumptions. The proposed method has smaller standard error compared to WB, which is consistent with the simulation results. Nevertheless, this difference is not as prominent compared to their difference from the logistic regression.

Table 4: Estimation results in the Pima Indians diabetes data

	LR			SAUC/WB			PROP		
Variable	Est	SE	z-score	Est	SE	z-score	Est	SE	z-score
Age	16	10	1.593	16	13	1.256	14	11	1.250
BP	-14	5	-2.540	-10	6	-1.718	-11	5	-2.331
BMI	94	16	5.945	72	26	2.808	65	14	4.670
DPF	986	312	3.160	992	7	137.1	992	4	248.5
Glucose	37	4	9.481	32	11	2.858	33	5	6.420
Insulin	-1	1	-1.322	-1	1	-0.807	-1	1	-1.192
PG	129	33	3.840	92	41	2.258	102	38	2.711
ST	1	7	0.090	-3	7	-0.375	0	6	0.018

BP: Blood pressure; BMI: Body mass index; DPF: Diabetes pedigree function; PG: Pregnancies; ST: Skin Thickness.
Est: estimated coefficient (×1000); SE: Standard error (×1000).
LR: logistic regression; SAUC/WB: Sigmoid-smoothed AUC used for coefficient estimation with weighted bootstrap used for variance estimation; PROP: porposed method.

4. Discussion

In this article, we have proposed a new empirical AUC optimization method for biomarker combinations. Our proposal has a number of distinctive and desirable features. First, we tackle the general biomarker combination problem without designating an anchor biomarker, as required by many existing methods. Second, we reformulate the norm-constrained empirical AUC as an unconstrained optimization problem, which facilitates both computation and statistical inference. Third, our estimator is defined in terms of a sequence of smoothed empirical AUC functions approaching the unsmoothed one, eliminating the need to specify a kernel bandwidth for smoothing —

a necessity in many current approaches. Finally, our inference procedure is sample-based and computationally efficient. The proposed optimization algorithm is shown to have competitive computational and statistical performance.

While we have focused on the standard empirical AUC maximization, our proposed methods immediately extend to several other problems. One such problem is the empirical maximization of center-adjusted AUC considered by Meisner et al. (2019), to address multicenter biomarker studies. Another problem of interest is the empirical maximization of the partial rank correlation estimate (Khan and Tamer, 2007) for censored survival outcomes. Though the type of outcome is different, the objective function is similar and would be amenable to our techniques.

Nevertheless, several issues warrant further investigation. One is the performance estimation of the resulting combination. It is well known that the maximized empirical AUC tends to be overly optimistic for prediction performance. Cross-validation is a standard approach to address the overoptimism, but can be computationally demanding (Huang et al., 2011). Our proposed computational techniques would be useful in this regard. Another issue is biomarker selection for combination. As high-dimensional biomarkers like genetic and microbiome data become more accessible, selecting relevant biomarkers becomes increasingly important. Lin et al. (2011) considered penalized empirical AUC maximization for simultaneous biomarker selection and combination estimation. Similar procedures may be developed to incorporate and adapt our optimization techniques developed herein. However, the statistical properties of these approaches require further examination.

Supplementary Material

The online Supplementary Material contains additional simulation results.

Acknowledgment

We sincerely thank the reviewers for their thoughtful comments and constructive suggestions, which have helped to enhance the quality of this manuscript. The authors were supported in part by NIH grants R01 CA230268 and P30 AI050409.

Appendix

A. Technical Proof

Regularity conditions adapted from Sherman (1993) are imposed.

Assumption 1. *The maximizer* $\boldsymbol{\beta}_0$ *of* $A(\boldsymbol{b})$ *over* $\mathcal{B} = \{\boldsymbol{b} : \|\boldsymbol{b}\|_2 = 1, \boldsymbol{b} \in \mathbb{R}^d\}$ *is unique.*

Assumption 2. *The biomarkers satisfy:*

1. *The support of* $\mathbf{X}$ *is not contained in any linear subspace of* $\mathbb{R}^d$.

2. *There exists one component* X_t *in* $\mathbf{X}$, $t \in \{1, ..., d\}$ *with its corresponding coefficient in* $\boldsymbol{\beta}_0$ *being non-zero, such that conditional on the other* $d-1$ *components,* X_t *has a everywhere positive density function with respect to the Lebesgue measure.*

Assumption 3. *Let* $\mathcal{N}$ *denote a neighborhood of* $\boldsymbol{\beta}_0$*, and* $\|\cdot\|$ *denote the matrix norm* $\|(a_{ij})\| = (\sum_{i,j} a_{ij}^2)^{1/2}$. *For each pair* $(y, \boldsymbol{x})$ *of possible values of* $(Y, \mathbf{X})$*,*

1. *the second derivative of* $\tau(y, \boldsymbol{x}; \boldsymbol{\beta})$ *with respect to* $\boldsymbol{\beta}$ *exist in* $\mathcal{N}$;

2. *there is an integrable function* $M(y, \boldsymbol{x})$ *such that for all* $\boldsymbol{\beta}$ *in* $\mathcal{N}$,

$$\|\nabla_2\tau(y, \boldsymbol{x}; \boldsymbol{\beta}) - \nabla_2\tau(y, \boldsymbol{x}; \boldsymbol{\beta}_0)\|_2 \leq M(y, \boldsymbol{x})\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2;$$

3. $\mathrm{E}\left(\|\nabla_1\tau(Y, \mathbf{X}; \boldsymbol{\beta}_0)\|_2\right)^2 < \infty$;

4. $\mathrm{E}\left(\|\nabla_2\tau(Y, \mathbf{X}; \boldsymbol{\beta}_0)\|_2\right) < \infty$;

5. *The matrix* $\mathbf{H}_p$ *is strictly negative definite.*

Assumption 1 and 2 are used to establish consistency (Han, 1987). Assumption 3 is a set of general conditions used to establish asymptotic normality of the estimator, similar to the assumptions in Sherman (1993).

A.1 Proof of Proposition 1

Proof. Similar to expansion (7) of Sherman (1993), we can show that

$$\begin{aligned}\widehat{A}(\boldsymbol{b}) - \widehat{A}(\boldsymbol{\beta}_0) = n^{-1/2}(\boldsymbol{b}-\boldsymbol{\beta}_0)^T\mathbf{W}_n + \frac{1}{2}(\boldsymbol{b}-\boldsymbol{\beta}_0)^T\mathbf{H}_0(\boldsymbol{b}-\boldsymbol{\beta}_0)\\ + o_p(\|\boldsymbol{b}-\boldsymbol{\beta}_0\|_2^2) + o_p(n^{-1}),\end{aligned} \tag{A.1}$$

where $\mathbf{W}_n = \{(N_1N_0)^{-1}(n-1)n\}n^{-1/2}\sum_i \nabla\tau_0(Y,\mathbf{X};\boldsymbol{\beta}_0)$. Because $N_1/n \xrightarrow{p} \rho$, $(N_1N_0)^{-1}(n-1)n \xrightarrow{p} \{\rho(1-\rho)\}^{-1}$. On the other hand,

$$n^{-1/2}\sum_{i=1}^{n}\nabla\tau_0(Y,\mathbf{X};\boldsymbol{\beta}_0) \xrightarrow{d} N(0, \mathrm{E}\{\nabla\tau_0(Y,\mathbf{X};\boldsymbol{\beta}_0)\}^{\otimes 2}).$$

Thus, by Slutsky's theorem, $\mathbf{W}_n \xrightarrow{d} N(\mathbf{0}, \mathbf{V}_0)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0195

Write $P(\boldsymbol{b}) = w(\|\boldsymbol{b}\|_2 - 1)^2$, it follows that

$$\{\widehat{A}(\boldsymbol{b}) - P(\boldsymbol{b})\} - \{\widehat{A}(\boldsymbol{\beta}_0) - P(\boldsymbol{\beta}_0)\} = n^{-1/2}(\boldsymbol{b} - \boldsymbol{\beta}_0)^T \mathbf{W}_n$$

$$+\frac{1}{2}(\boldsymbol{b} - \boldsymbol{\beta}_0)^T (\mathbf{H}_0 - 2w\boldsymbol{\beta}_0^{\otimes 2})(\boldsymbol{b} - \boldsymbol{\beta}_0) + o_p(\|\boldsymbol{b} - \boldsymbol{\beta}_0\|_2^2 + n^{-1}) \text{(A.2)}$$

Note that the rank of $\mathbf{H}_0$ is $d-1$, and $\boldsymbol{\beta}_0$ is not in the span of the column space of $\mathbf{H}_0$ due to the norm constraint. Therefore, the Hessian matrix $\mathbf{H}_P$ for the new objective function is invertible. Following Theorem 2 in Sherman (1993), we have

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \mathbf{H}_P^{-1}\mathbf{W}_n + o_p(1) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_0), \tag{A.3}$$

where $\boldsymbol{\Sigma}_0 = \mathbf{H}_P^{-1}\mathbf{V}_0\mathbf{H}_P^{-1}$. □

A.2 Proof of Theorem 1

Proof. First, we show that $\widetilde{\boldsymbol{\beta}}_\sigma$ is strongly consistent. By definition, $\widehat{A}(\boldsymbol{b})$ is bounded by 1 for any $\boldsymbol{b} \in \mathcal{B}$. Han (1987) showed that

$$\lim_n \sup_{\boldsymbol{b}\in\mathcal{B}} |\widehat{A}(\boldsymbol{b}) - A(\boldsymbol{b})| = 0, \quad a.s. \tag{A.4}$$

Following a similar argument, we can show that such uniform convergence holds for $\widetilde{A}_{s^{-1}\sigma(\boldsymbol{b})}$ with $\sigma = o(n^{-1/2})$:

$$\lim_{n} \sup_{\boldsymbol{b}\in\mathcal{B}} |\widetilde{A}_{s^{-1}\sigma}(\boldsymbol{b}) - A(\boldsymbol{b})| = 0, \quad a.s. \tag{A.5}$$

With Assumption 1, by uniform convergence over $\mathcal{B}$, we further have $\widetilde{\boldsymbol{\beta}}_\sigma \rightarrow \boldsymbol{\beta}_0$ almost surely as $n \rightarrow \infty$.

Next, we establish the asymptotic normality of $\widetilde{\boldsymbol{\beta}}_\sigma$. We first introduce an expansion for a U-statistic defined as:

$$U(\boldsymbol{b}, \delta) = \frac{1}{n(n-1)} \sum_{i\neq j} I(Y_i > Y_j) I(\boldsymbol{b}^T\mathbf{X}_i - \boldsymbol{b}^T\mathbf{X}_j > \delta), \tag{A.6}$$

where $\delta \in \mathbb{R}$. The kernel of $U(\boldsymbol{b}, \delta)$ is:

$$\tau(y, \boldsymbol{x}; \boldsymbol{b}, \delta) = \mathrm{E}\left[I(y > Y) I\{(\boldsymbol{x} - \mathbf{X})^T\boldsymbol{b} > \delta\} + I(y < Y) I\{(\boldsymbol{x} - \mathbf{X})^T\boldsymbol{b} < \delta\} \right].$$

Following the arguments in Sherman (1993), we can show that, uniformly over any $o_p(1)$ neighbourhood of $(\boldsymbol{\beta}_0, 0)$,

$$\begin{aligned} U(\boldsymbol{b}, \delta) - U(\boldsymbol{\beta}_0, 0) &= n^{-1/2}\{(\boldsymbol{b}^T, \delta) - (\boldsymbol{\beta}_0^T, 0)\}\mathbf{W}_n^{(e)} \\ &\quad + \frac{1}{2}\{(\boldsymbol{b}^T, \delta) - (\boldsymbol{\beta}_0^T, 0)\}\mathbf{H}\{(\boldsymbol{b}^T, \delta) - (\boldsymbol{\beta}_0^T, 0)\}^T \\ &\quad + o_p\{\|(\boldsymbol{b}^T, \delta) - (\boldsymbol{\beta}_0^T, 0)\|_2^2\} + o_p(n^{-1}), \qquad \text{(A.7)} \end{aligned}$$

where $\mathbf{W}_n^{(e)} = n^{-1/2}\sum_i \nabla_1 \tau(Y_i, \mathbf{X}_i; \boldsymbol{\beta}_0, 0)$, and $2\mathbf{H} = \mathrm{E}\{\nabla_2 \tau(Y, \mathbf{X}; \boldsymbol{\beta}_0, 0)\}$.

Let $\widetilde{\Gamma}_n(\boldsymbol{b}) = \mathrm{E}_Z\{U(\boldsymbol{b}, Z) - U(\boldsymbol{\beta}_0, 0)\}$, where Z is an independent random variable with c.d.f and E_Z is taking expectation with respect to Z. $(g_\sigma - f_\sigma)(\cdot)$. Notice that $Z = o_p(n^{-1/2})$ since σ is $o_(n^{-1/2})$ and the support of Z is $[-\sigma, \sigma]$. Thus, by expansion (A.7), uniformly over any $o_p(1)$ neighbourhood of $\boldsymbol{\beta}_0$, we have

$$\begin{aligned} \widetilde{\Gamma}_n(\boldsymbol{b}) &= n^{-1/2}\{\boldsymbol{b}^T, \mathrm{E}(Z)\}\mathbf{W}_n^{(e)} + \mathrm{E}\{(\boldsymbol{b}^T, Z)\mathbf{H}(\boldsymbol{b}^T, Z)^T\}/2 \\ &\quad + o_p(\|\boldsymbol{b}\|_2^2) + o_p(n^{-1}) \\ &= n^{-1/2}\boldsymbol{b}^T\mathbf{W}_n + \boldsymbol{b}^T\mathbf{H}_0\boldsymbol{b}/2 + \boldsymbol{b}_p\mathrm{E}(Z^2)/2 \\ &\quad + o_p(\|\boldsymbol{b}\|_2^2) + o_p(n^{-1}), \qquad \text{(A.8)} \end{aligned}$$

where $2\mathbf{H}_0 = \mathrm{E}\{\partial^2\tau(Y, \mathbf{X}; \boldsymbol{\beta}_0, 0)/\partial \boldsymbol{b}^2\}$, $2\boldsymbol{b}_p = \mathrm{E}\{\partial^2\tau(Y, \mathbf{X}; \boldsymbol{\beta}_0, 0)/\partial \boldsymbol{b}\partial\delta\}$, and $\mathbf{W}_n$ is the first $d-1$ elements of $\mathbf{W}_n^{(e)}$. The second equality follows

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0195

from the symmetric distribution of Z. Replacing $\boldsymbol{b}$ with $\boldsymbol{\beta}_0$ in Equation (A.8) and subtracting it from $\widetilde{\Gamma}_n(\boldsymbol{b})$, we have

$$\begin{aligned}\widetilde{\Gamma}_n(\boldsymbol{b}) - \widetilde{\Gamma}_n(\boldsymbol{\beta}_0) &= n^{-1/2}(\boldsymbol{b}-\boldsymbol{\beta}_0)^T\mathbf{W}_n + \frac{1}{2}(\boldsymbol{b}-\boldsymbol{\beta}_0)^T\mathbf{H}_0(\boldsymbol{b}-\boldsymbol{\beta}_0)\\ &\quad + o_p(\|\boldsymbol{b}-\boldsymbol{\beta}_0\|_2^2) + o_p(n^{-1}). \qquad \text{(A.9)}\end{aligned}$$

Notice that expansion (A.9) is very similar to expansion (A.1). Following similar arguments in the proof of Proposition 1, we have $\sqrt{n}(\widetilde{\boldsymbol{\beta}}_\sigma - \widehat{\boldsymbol{\beta}}) = o_p(1)$. □

A.3 Proof of Theorem 2

Proof. To establish consistency for the covariance matrix, we only need to show

$$\widehat{\mathbf{H}}_P(\widehat{\boldsymbol{\beta}}, \boldsymbol{\Sigma}) \xrightarrow{p} \mathbf{H}_0 + 2w\boldsymbol{\beta}_0^{\otimes 2} \qquad \text{(A.10)}$$

and

$$\widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}}, \boldsymbol{\Sigma}) \xrightarrow{p} \mathbf{V}_0 \qquad \text{(A.11)}$$

for any positive definite matrix $\boldsymbol{\Sigma}$. Without loss of generality, let $\boldsymbol{\Sigma} = \mathbf{I}$. Denote

$$Q(\widehat{\boldsymbol{\beta}}) = \frac{1}{N_1 N_0}\sum_{i=1}^{N_1}\sum_{j=1}^{N_0}\Phi\left(\frac{\sqrt{n}\mathbf{X}_{ij}^T\widehat{\boldsymbol{\beta}}}{\sigma_{ij}}\right).$$

It also has the integral representation

$$\int Q(\widehat{\boldsymbol{\beta}} + \boldsymbol{z}/\sqrt{n})(2\pi)^{-d/2}\exp\bigl(-\|\boldsymbol{z}\|_2^2/2\bigr)d\boldsymbol{z}. \tag{A.12}$$

Therefore, the same arguments to establish consistency in the proof for Theorem 2 in Zhang et al. (2018), combined with Expansion (A.1) applies to show

$$\nabla_2 Q(\widehat{\boldsymbol{\beta}}) \xrightarrow{p} \mathbf{H}_0 \tag{A.13}$$

and (A.11). In addition,

$$\nabla_2 P(\widehat{\boldsymbol{\beta}}) = 2w\left(\frac{\widehat{\boldsymbol{\beta}}^{\otimes 2}}{\|\widehat{\boldsymbol{\beta}}\|_2^3} + \frac{\|\widehat{\boldsymbol{\beta}}\|_2 - 1}{\|\widehat{\boldsymbol{\beta}}\|_2}\mathbf{I}\right). \tag{A.14}$$

Because $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$, by continuous mapping theorem, we have $\nabla_2 P(\widehat{\boldsymbol{\beta}}) \xrightarrow{p} 2w\boldsymbol{\beta}_0^{\otimes 2}$. Combining (A.14) with (A.13), we have (A.10).

□

References

Bélisle, C. J. (1992). Convergence theorems for a class of simulated annealing algorithms on $\mathbb{R}^d$. *Journal of Applied Probability 29*(4), 885–895.

Broyden, C. G. (1970). The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics 6*(1), 76–90.

Chen, X., A. Vexler, and M. Markatou (2015). Empirical likelihood ratio confidence interval estimation of best linear combinations of biomarkers. *Computational Statistics & Data Analysis 82*, 186–198.

Fletcher, R. (1970). A new approach to variable metric algorithms. *The computer journal 13*(3), 317–322.

Fong, Y., S. Yin, and Y. Huang (2016). Combining biomarkers linearly and nonlinearly for classification using the area under the roc curve. *Statistics in medicine 35*(21), 3792–3809.

Goldfarb, D. (1970). A family of variable-metric methods derived by variational means. *Mathematics of computation 24*(109), 23–26.

Han, A. K. (1987). Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. *Journal of Econometrics 35*(2-3), 303–316.

Huang, X., G. Qin, and Y. Fang (2011). Optimal combinations of diagnostic tests based on auc. *Biometrics 67*(2), 568–576.

Huang, Y. and M. G. Sanda (2022). Linear biomarker combination for constrained classification. *Ann. Statist. 50*(5), 2793–2815.

Khan, S. and E. Tamer (2007). Partial rank estimation of duration models with general forms of censoring. *Journal of Econometrics 136*(1), 251–280.

Lin, H., L. Zhou, H. Peng, and X.-H. Zhou (2011). Selection and combination of biomarkers using roc method for disease classification and prediction. *Canadian Journal of Statistics 39*(2), 324–343.

Ma, S. and J. Huang (2007). Combining multiple markers for classification using roc. *Biometrics 63*(3), 751–757.

Meisner, A., C. R. Parikh, and K. F. Kerr (2019). Biomarker combinations for diagnosis and prognosis in multicenter studies: Principles and methods. *Statistical methods in medical research 28*(4), 969–985.

Pepe, M. S., T. Cai, and G. Longton (2006). Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics 62*(1), 221–229.

Pepe, M. S. and M. L. Thompson (2000). Combining diagnostic test results to increase accuracy. *Biostatistics 1*(2), 123–140.

Shanno, D. F. (1970). Conditioning of quasi-newton methods for function minimization. *Mathematics of computation 24*(111), 647–656.

Sherman, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica: Journal of the Econometric Society 61*(1), 123–137.

Smith, J. W., J. E. Everhart, W. Dickson, W. C. Knowler, and R. S. Johannes (1988). Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the annual symposium on computer application in medical care*, pp. 261. American Medical Informatics Association.

Vexler, A., A. Liu, E. F. Schisterman, and C. Wu (2006). Note on distribution-free estimation of maximum linear separation of two multivariate distributions. *Nonparametric Statistics 18*(2), 145–158.

Yuille, A. L. and A. Rangarajan (2003). The concave-convex procedure. *Neural computation 15*(4), 915–936.

Zhang, J., Z. Jin, Y. Shao, and Z. Ying (2018). Statistical inference on transformation models: a self-induced smoothing approach. *Journal of Nonparametric Statistics 30*(2), 308–331.

Department of Biostatistics and Bioinformatics, Emory University, Atlanta, Georgia, U.S.A

E-mail: yuxuan.chen@emory.edu

Department of Biostatistics and Bioinformatics, Emory University, Atlanta, Georgia, U.S.A

E-mail: yhuang5@emory.edu