Nonparametric Estimation of the AUC of an Index with Estimated Parameters

Alexis Doyle-Connolly, Haben Michael
University of Massachusetts

Abstract. We describe a nonparametric method of estimating the AUC of an index $\beta^T x$ when β is estimated from the same data, with a focus on nonparametric estimation of the difference of the AUCs of two distinct indices. The difference in AUCs is a popular method for comparing AUCs that has come under scrutiny due to the frequent misapplication of a standard inferential method. Besides providing a valid inferential method for the case where the difference of AUCs is nonzero, we enumerate many common situations in which the standard method is applicable. We show by simulation that there is no great loss in efficiency in using our more generally applicable estimator when the standard method is applicable, while invalid application of the standard method may lead either to loss of Type I or II error control. As an application, we apply the proposed method to re-analyze indices based on the Framingham Heart Study.

1 Introduction

The AUC is a measure of how effectively a marker discriminates between two classes, and the difference in AUCs compares the discrimination of two markers. In the medical sciences, the marker is often a linear combination β of a set of subject characteristics x. We refer here to $\beta^T x$ as an "index" and its AUC as an "index AUC." In medical fields, comparison of markers often takes the form of comparing two sets of patient characteristics x and y, with indexes $\beta^T x, \gamma^T y$. The characteristics are often nested, $x \subset y$, as when investigating the impact on discrimination of additional factors $y \backslash x$. The difference in AUCs has been described by experts as one of the most widely used measures of the difference in discrimination (Demler et al., 2017).

A related way of measuring the impact of additional covariates is to directly compare the coefficients β and γ under models parameterized by β and γ. The result of this comparison may conflict with the comparison of AUCs. A series of papers in the 2010s noted the "baffling" and "perplexing" contradictions between the two methods, calling into question the validity of the difference in AUCs when evaluating markers (Seshan et al., 2013; Tzoulaki et al., 2009). While the source of the contradiction was soon identified (Demler et al., 2012, 2017; Seshan et al., 2013; Heller et al., 2017), remedies have been slower to arrive.We present here a partial remedy, a method to nonparametrically estimate the difference in AUCs under the assumption that it is nonzero.

The remainder of the paper is organized as follows. Next we give more background on the problem of inference on the difference of index AUCs and summarize approaches available in the literature. In Section 3 we derive a nonparametric estimator for the difference of index AUCs, along the way indicating the derivation of the standard estimator. In Section 4 we briefly discuss some practical issues in applying the proposed estimator. In Section 5 we discuss a range of examples both those when the standard estimator is valid and those when the proposed estimator is needed, both for the index AUC itself and the difference in index AUCs. In Section 6 we examine the finite-sample performance of the proposed estimator and competing estimators. in Section 7 we apply the proposed estimator to re-analyze the Framingham Heart Study data considered by Demler et al. (2011, 2017). We conclude and suggest extensions and directions for future works in Section 8. Software implementing the proposed estimator and the routines used in the simulation and data analysis sections are publicly available at the second author's website.

2 Background

An observation is modeled as a pair consisting of covariates W and a binary status indicator D,

$$(W, D), W \in \mathbb{R}^p, P(D = 0) = 1 - P(D = 1) \in (0, 1). \tag{1}$$

Denote by $X_0 \sim F, X_1 \sim G$ the RVs and distributions obtained by conditioning W on $D = 0$ and $D = 1$. We use "control" and "case" generically to refer to these conditional RVs and distributions. Let $(W_1, D_1), \dots, (W_{M+N}, D_{M+N})$, be an IID sample under (1), with the class variables

$$X_{01}, \dots, X_{0M} \overset{IID}{\sim} F, X_{11}, \dots, X_{1N} \overset{IID}{\sim} G, M = \sum\{D = 0\}, N = \sum\{D = 1\}.$$

Vectors $\hat{\beta} \in \mathbb{R}^p$ and $\hat{\gamma} \in \mathbb{R}^p$ are obtained based on the sample by some procedure such as logistic regression. They are assumed to have finite probability limits β^* and γ^* as $M, N \to \infty$ under this procedure.

The AUC, measuring how effectively a scalar marker discriminates between two classes, is the probability the marker associated with one class is less than a stochastically independent marker associated with the other class, with ties weighted by half. A nonparametric estimator of the AUC is the sample proportion of markers in one class less than the markers in the other, with ties weighted by half. In the case that the markers are indexes with estimated coefficient $\hat{\beta}$, the estimator is

$$\hat{\theta} = \frac{1}{MN} \sum_{i,j} \psi(\hat{\beta}^T X_{0i}, \hat{\beta}^T X_{1j}), \tag{2}$$

where $\psi : (u, v) \mapsto \{u < v\} + \frac{1}{2}\{u = v\}$. The difference of index AUCs is estimated nonparametrically by

$$\Delta\hat{\theta} = \frac{1}{MN} \sum_{i,j} \psi(\hat{\beta}^T X_{0i}, \hat{\beta}^T X_{1j}) - \frac{1}{MN} \sum_{i,j} \psi(\hat{\gamma}^T X_{0i}, \hat{\gamma}^T X_{1j}). \tag{3}$$

In applied settings, an explicit probabilty model may not be specified, and often the estimation methods for $\hat{\beta}$ and $\hat{\gamma}$ imply inconsistent models (see Example 6). Nevertheless inference is sought (see Seshan et al. (2013) for enumerations of the many applications), particularly 1) whether the difference in the AUCs of the two markers $\hat{\beta}^T x$ and $\hat{\gamma}^T x$ is in some limiting sense nonzero, and if so, 2) the magnitude of the difference. Assume here that limiting sense is the difference of AUCs of the indexes at the starred parameters, so the target of inference is

$$\Delta\theta = P({\beta^*}^T X_{0i} < {\beta^*}^T X_{1j}) - P({\gamma^*}^T X_{0i} < {\gamma^*}^T X_{1j}). \tag{4}$$

The statistic (3) may be viewed as a U-statistic process with two-sample kernel $(x, y) \mapsto \psi(\beta^T x, \beta^T y) - \psi(\gamma^T x, \gamma^T y)$, indexed by β, γ, and evaluated at the random vectors $\hat{\beta}, \hat{\gamma}$. This statistic presents two complications for an analysis using basic U-statistics theory.

1. Under the null of no difference, $\Delta\theta = 0$, the statistic (4) is often a degenerate U-statistic. The asymptotic distribution of a degenerate U-statistic is a weighted combination of chi-squares, with weights depending on the distributions of the observations. In the case of the difference of index AUCs, estimators have been presented only in a few specific cases, e.g., Heller et al. (2017), and the null distribution for common coefficient estimation methods such as logistic regression remains "intractable" (Lee, 2021).

Instead, the literature has proposed the use of more convenient testing problems equivalent in certain settings to testing $\Delta\theta = 0$. Demler et al. (2011) show that when the covariates are Gaussian and the coefficient estimation procedure is LDA, the null is the same as testing for equality of the Mahalonobis distance between the two class distrbutions. An F-test, valid in finite samples, may therefore be used instead of testing the AUC directly. Pepe et al. (2013) describe a more general approach. The risk function for a binary RV D based on a set of covariates $W \in \mathbb{R}^p$, $\rho_W : \mathbb{R}^p \to \mathbb{R}$, is the function $w \mapsto P(D = 1 \mid W = w)$. Let $(D_0, W_0, W_0'), (D_1, W_1, W_1')$ be IID. The authors show that the null of equal AUCs of the risks,

$$\begin{aligned} &P(\rho_{W,W'}(W_0, W_0') < \rho_{W,W'}(W_1, W_1') \mid D_0 = 0, D_1 = 1) \\ &= P(\rho_W(W_0) < \rho_W(W_1) \mid D_0 = 0, D_1 = 1) \end{aligned}$$

holds if and only if the risk functions are equal, $\rho_{W,W'} = \rho_W$. Often the coefficient estimation procedure is of secondary importance and the goal of testing the null $\Delta\theta = 0$ is to test if certain additional covariates improve discrimination. In this case, the result shows that the test may be based on the risks instead. Even if interest lies in testing for the difference in AUCs where $\hat{\beta}, \hat{\gamma}$ are obtained through a particular estimation procedure, e.g., logistic regression, for many estimation procedures there is a monotone link connecting the limiting index to the risk, e.g., the expit function (see Example 3.1). Since the AUC is invariant to monotone transformations, the risk may still be used to test for a difference.

A drawback to this approach is it requires knowing the true risk function. If the null distribution of $\Delta\hat{\theta}$ were available, one might directly compare the discrimination of the indices $\hat{\beta}^T W$ and $\hat{\gamma}^T W$, and possibly use the indices in practice, without knowing the correct risk function. However, unless computing the null distribution of the $\Delta\hat{\theta}$ calls for fewer modeling assumptions, improved efficiency, or some other advantage, one may as well test risk functions.

The null distribution of $\Delta\hat{\theta}$ is analyzed in Michael et al. (2023) and we consider the alternative case $\Delta\theta \neq 0$ in the remainder. Experts currently recommend first testing for equality of risks, as just described, and producing confidence intervals for $\Delta\theta$ only upon a finding that $\Delta\theta \neq 0$ (e.g., Demler et al. (2017)). The focus in this paper on the alternative $\Delta\theta \neq 0$ has the goal of enabling practitioners to carry out this recommended course.

2. A second issue is that $\hat{\beta}, \hat{\gamma}$ are estimated from the data, so that the observations on which the statistic (3) is based are not IID. Non-degenerate U-statistics with estimated parameters are typically still normal though estimation of the parameter may affect the asymptotic distribution. This issue is addressed in the remainder.

3 Theory

The usual approach to finding the asymptotic distribution of a non-degenerate U-statistic, which we adopt, is to find an asymptotically equivalent IID average to which the CLT can be applied. The asymptotic variance may then be estimated using the sample variance of the terms of the IID average.

For control and case distributions F, G on $\mathbb{R}^p$, a control observation $X \sim F$ and independent case observation $Y \sim G$, and a vector β, denote the AUC of the index $P(\beta^T X < \beta^T Y)$ as

$$\theta(F, G, \beta) = \int \psi(\beta^T x, \beta^T y) dF(x) dG(y).$$

where $\psi(u, v) = \{u < v\} + \frac{1}{2}\{u = v\}$ for real u, v. With this notation, $\Delta\hat{\theta} = \theta(\hat{F}, \hat{G}, \hat{\beta}) - \theta(\hat{F}, \hat{G}, \hat{\gamma})$. We write each of the two terms of the difference as an IID average, and later take the difference to represent $\Delta\hat{\theta}$ as an IID average. Decompose the centered estimate $\theta(\hat{F}, \hat{G}, \hat{\beta}) - \theta(F, G, \beta^*)$ as a sum of two terms, reflecting the two sources of estimation, the AUC estimation and the coefficient estimation,

$$\begin{aligned} &\theta(\hat{F}, \hat{G}, \hat{\beta}) - \theta(F, G, \beta^*) \\ &= \theta(F + \delta F, G + \delta G, \beta^* + \delta\beta) - \theta(F, G, \beta^* + \delta\beta) && (5) \\ &+ \theta(F, G, \beta^* + \delta\beta) - \theta(F, G, \beta^*) && (6) \end{aligned}$$

where $\delta F = \hat{F} - F$, etc. We give asymptotically linear representations of terms (5) and (6).

Term (5): As the function $\theta(\cdot, \cdot, \beta)$ is bilinear for any β,

$$\begin{aligned} &\theta(F + \delta F, G + \delta G, \beta^* + \delta\beta) - \theta(F, G, \beta^* + \delta\beta) \\ &= \theta(\delta F, G, \beta^* + \delta\beta) + \theta(F, \delta G, \beta^* + \delta\beta) + \theta(\delta F, \delta G, \beta^* + \delta\beta). && (7) \end{aligned}$$

The third and final term in (7),

$$\theta(\delta F, \delta G, \beta^* + \delta\beta) = \int (\psi(\hat{\beta}^T x, \hat{\beta}^T y)) d(\hat{F} - F)(x) d(\hat{G} - G)(y),$$

is $o((M + N)^{-1/2})$. Heuristically, the integrand is nonnegative and bounded by 1, while the integrators $\hat{F} - F, \hat{G} - G$ are each $O((M + N)^{-1/2})$. For details see Michael et al. (2023), Lemma 1, showing that the term is in fact $O((M+N)^{-1})$ uniformly in a neighborhood of betastar.

Conditionally on any sequence of statuses $D_1, D_2, \ldots$, for fixed $\beta^*+\delta\beta$, the first two terms in (7) are centered IID averages. That the randomness in $\delta\beta$ is asymptotically negligible at the $\sqrt{M+N}$ rate,

$$\theta(\delta F, G, \beta^*+\delta\beta) + \theta(F, \delta G, \beta^*+\delta\beta) = \theta(\delta F, G, \beta^*) + \theta(F, \delta G, \beta^*) + o((M+N)^{-1/2}),$$

follows from empirical process theory, in particular "stochastic equicontinuity" of the empirical processes $\beta \mapsto \theta(\delta F, G, \beta), \beta \mapsto \theta(F, \delta G, \beta)$ Details of this calculation are carried out in e.g., Sherman (1993), Sec. 5.

Therefore,

$$\begin{aligned}
&\theta(F+\delta F, G+\delta G, \beta^*+\delta\beta) - \theta(F, G, \beta^*+\delta\beta) \\
&= -\frac{1}{M}\sum_{i=1}^{M}(1 - G(\beta^{*T}X_{0i}) - \theta(F,G,\beta^*)) + \frac{1}{N}\sum_{i=1}^{N}(F(\beta^{*T}X_{1i}) - \theta(F,G,\beta^*)) + o((M+N)^{-1/2}) \qquad (8) \\
&= \frac{1}{M+N}\sum_{i=1}^{M+N}\left(-\frac{\{D_i=0\}}{P(D=0)}(1 - G(\beta^{*T}W_i) - \theta(F,G,\beta^*)) + \frac{\{D_i=1\}}{P(D=1)}(F(\beta^{*T}W_i) - \theta(F,G,\beta^*))\right) \\
&+ o((M+N)^{-1/2}) \qquad (9)
\end{aligned}$$

This IID representation is known as the Hoeffding decomposition of a U-statistic (Hoeffding, 1948). It may be helpful in this context to view it as the first von Mises derivative, a type of functional derivative, so that $\Delta\hat{\theta} - \Delta\theta$ is resolved as the sum of this functional derivative, corresponding to (5), is resolved by a functional Taylor expansion and a usual finite-dimensional Taylor expansion, discussed next, corresponding to (6). The above derivation of the Hoeffding decomposition differs from the usual one to allow for random $\hat{\beta}$, and (9) shows that this randomness is asymptotically negligible, at least as far as term (5) is concerned. In situations where term (6) is negligible, e.g., if $\hat{\beta} = \beta$ were known rather than estimated (see Section 5 for additional scenarios when (6) is negligible), the CLT may be applied to get the asymptotic distribution of $\Delta\hat{\theta}$. The approach of DeLong et al. (1988) in such situations is to estimate F, G in (8) using the empirical CDFs $\hat{F}, \hat{G}$,

$$-\frac{1}{M}\sum_{i=1}^{M}(1 - \hat{G}(\beta^T X_{0i}) - \theta(\hat{F}, \hat{G}, \beta)) + \frac{1}{N}\sum_{i=1}^{N}(\hat{F}(\beta^T X_{1i}) - \theta(\hat{F}, \hat{G}, \beta)),$$

and simply use the sample variance of the terms for inference.

Term (6): Assume $\sqrt{n}(\hat{\beta} - \beta^*) \to 0$ in probability, $\beta \mapsto \theta(F, G, \beta)$ is differentiable at β^*. Let the function $\psi_{\hat{\beta}}$ represent the estimator $\hat{\beta}$ as an IID mean

$$\hat{\beta} - \beta^* = (M+N)^{-1}\sum_{i=1}^{M+N}\psi_{\hat{\beta}}(W_i) + o((M+N)^{-1/2})$$

i.e., $\psi_{\hat{\beta}}$ is an influence function for $\hat{\beta}$. Then (6) is

$$\begin{aligned}
&\theta(F, G, \beta + \delta\beta) - \theta(F, G, \beta) \\
&= (\hat{\beta} - \beta^*)\frac{\partial}{\partial\beta}\theta(F, G, \beta) + o_P((M+N)^{-1/2}) \\
&= \frac{\partial}{\partial\beta}\theta(F, G, \beta)(M+N)^{-1}\sum_{i=1}^{M+N}\psi_{\hat{\beta}}(W_i) + o_P((M+N)^{-1/2}).
\end{aligned}$$

Putting the two parts together,

$$\begin{aligned}
&\theta(\hat{F}, \hat{G}, \hat{\beta}) - \theta(F, G, \beta) \\
&= \frac{1}{M+N}\sum_{i=1}^{M+N}\left(-\frac{\{D_i = 0\}}{P(D=0)}(1 - G(\beta^{*T}W_i) - \theta(F, G, \beta^*)) + \frac{\{D_i = 1\}}{P(D=1)}(F(\beta^{*T}W_i) - \theta(F, G, \beta^*))\right) \\
&+ \frac{\partial}{\partial\beta}\theta(F, G, \beta)\frac{1}{M+N}\sum_{i=1}^{M+N}\psi_{\hat{\beta}}(W_i) + o_P((M+N)^{-1/2}). \qquad (10)
\end{aligned}$$

Proposition 1. *Given $(W_1, D_1), \ldots, (W_{M+N}, D_{M+N})$ sampled under (1), and estimator $\hat{\beta}$ based on the sample, assume*

1. *an influence function for $\hat{\beta}$ is available,*
2. *$P(D = 0) \in (0, 1)$,*
3. *$\theta(F, G, \cdot)$ is differentiable at β^*.*

Then $(M+N)^{-1/2}(\theta(\hat{F}, \hat{G}, \hat{\beta}) - \theta(F, G, \beta^*))$ is asymptotically normal with mean zero and variance given by the variance of a term in (10)*. This variance may be consistently estimated as $\sqrt{M+N}$ times the sample variance of the terms in* (10)*.*

Take the difference with the same representation of another estimator, $\theta(\hat{F}, \hat{G}, \hat{\gamma})$, to obtain an IID representation of $\Delta\hat{\theta}$.

Corollary 2. *Given $(W_1, D_1), \ldots, (W_{M+N}, D_{M+N})$ sampled under (1), and estimators $\hat{\beta}, \hat{\gamma}$ based on the sample, assume*

1. *the assumptions of Proposition 1 apply to $\hat{\beta}, \hat{\gamma}$ each,*
2. *$\beta^* \neq \gamma^*$.*

Then $(M+N)^{-1}(\Delta\hat{\theta} - \Delta\theta)$ is asymptotically normal with mean zero and variance given by the variance of a term in the difference of IID means (10)*. This variance may be consistently estimated as $\sqrt{M+N}$ times the sample variance of the difference of terms as in* (10)

What goes wrong under the null? If $\beta^* = \gamma^*$ then the main terms in the Hoeffding-type decomposition (9) cancel when the difference is taken, leaving a term of order $o((M+N)^{-1/2})$. Moreover the derivatives in (6) are the same. In many situations where the index is derived from a well-specified model the derivative is 0 for at least one of the two AUCs being differenced. In that case (5) will also be $o((M+N)^{-1/2})$, and $\Delta\hat{\theta}$ will degenerate under the usual $\sqrt{M+N}$ normalization. The condition $\beta^* \neq \gamma^*$ is just sufficient. It is possible that in, e.g., a nested logistic model both full and reduced are misspecified, the Hoeffding term degenerates, the derivative vanishes at neither β^* nor γ^*, the influence functions of $\hat{\beta}$ and $\hat{\gamma}$ differ, and then the limit of $\Delta\hat{\theta}$ is still normal.

4 Estimation

Using Proposition 1 for inference usually requires nuisance parameters to be estimated: The CDFs F, G, the influence function, and the derivative term in (10). As mentioned in Section 3, even if the coefficient β were not estimated, the terms in the estimator corresponding to the Delong statistic would still require estimating the CDFs. The influence function and derivative term belong to the adjustment term (6) and represent an additional cost due to estimating the coefficients. Substituting consistent estimated parameters into the IID average is usually asymptotically negligible as long as the dependence is continuous, though the efficiency of the convergence may be affected. Nonparametric estimation of the derivative term, in particular, is often slower than the usual parametric rate as it requires estimating a random function in an interval.

Since the index AUC is invariant to changes to the parameter in the radial direction, $P(\beta^T X_0 < \beta^T X_1) = P(c\beta^T X_0 < c\beta^T X_1), c > 0$, it seems β is more naturally parameterized using spherical coordinates. This transformation would reduce by 1 the dimension of the nuisance parameters such as the derivative term. However, this transformation does not seem to be used in practice and we follow convention inusing linear coordinates.

Proposition 1 does not require that $\hat{\beta}$ or $\hat{\gamma}$ be estimated by a correctly specified model, only that they have some probability limit at the parametric rate. Though the procedure for obtaining the estimate $\hat{\beta}$ or $\hat{\gamma}$ and the associated influence function ψ often involve some parametric assumptions, we still term the procedure described here as "nonparametric" since the estimator in Proposition 1 is valid under misspecification of the coefficient model. Whatever the estimation procedure is it will be known to the analyst, so that an influence function may be chosen, if one exists.

5 Examples

Section 3 decomposes the statistics $\hat{\theta}$ and $\Delta\hat{\theta}$ as a sum of two terms (5), (6). The first corresponds to the estimation of the AUC by a U-statistic. The second is an adjustment term corresponding to the use of estimated coefficients. At times the adjustment vanishes in the limit, and the estimation of the coefficient may be ignored. In this case the usual Mann-Whitney U-statistic, in the case of θ, or the Delong statistic, in the case of $\Delta\theta$, may be used for asymptotic inference, provided of course the AUCs are distinct, as discussed in

Section 3. Whether or not the adjustment term vanishes, Proposition 1 may be used for asymptotic inference.

We give examples of data and models where coefficient estimation may and may not be ignored when carrying out asymptotic inference on the index AUC.

5.1 No effect of coefficient estimation

In the ordinary course, the coefficient estimation can be ignored in computing the index of a smooth AUC iff its derivative is 0 at the probability limit of the coefficient, w in the notation of Section 3. For the difference of two AUCs, the derivative of each must usually be 0 at the respective coefficient probability limits, $\frac{d}{d\beta}\theta(F,G,\beta)|_{\beta=\beta^*} = \frac{d}{d\beta}\theta(F,G,\beta)|_{\beta=\gamma^*} = 0$.

5.1.1 AUC

We first give examples where the coefficient estimation may be ignored in estimating the AUC of an index.

Example 1 (Estimator: MRC, covariate restrictions: none/nonparametric)**.** The maximum rank correlation method of computing the coefficients is

$$\hat{\beta} = \arg\max_{\beta:|\beta|=1} \theta(\hat{F}, \hat{G}, \beta).$$

The method is nonparametric. By construction the empirical AUC is stationary at the coefficient estimates, and under mild regularity conditions the AUC $\theta(F,G,\beta)$ is stationary at the probability limit β^*, as well (Sherman, 1993).

While the MRC is a nonparametric maximizer of $\beta \mapsto \theta(F,G,\beta)$, it may also happen under parametric models that the derivative vanishes at the probability limit of the coefficient vector. The following proposition, highlighted by McIntosh and Pepe (2002); Pepe et al. (2013), furnishes a class of examples. For two real functions of W, f_1 and f_2, let the relation $f_1 \sim_{(W,D)} f_2$ hold iff $f_1(W)$ has the same conditional distribution given D as a strictly increasing function of $f_2(W)$, i.e., there is a strictly increasing function $h : \mathbb{R} \to \mathbb{R}$ such that $P(f_1(W) < w|D = i) = P(h \circ f_2(W) < w|D = i)$ for all $w \in \mathbb{R}$ and $i = 0, 1$.

Proposition 3. *Given a random vector (W, D), W continuous, D binary, with index AUC $\theta(F, G, \cdot)$. Then,*

1. *The ROC curve of classifying D based on a real function of W is maximized pointwise by the likelihood ratio $w \mapsto f_{W|D=1}(w)/f_{W|D=0}(w)$, equivalently, the risk of D based on W, $\rho_W(\cdot)$.*

2. *The AUC of a real function f of W is maximal iff $f \sim_{(W,D)} \rho_W$.*

3. *Given a coefficient estimate $\hat{\beta}$ with probability limit β^*, if $\theta(F,G,\cdot)$ is differentiable at β^*, and ${\beta^*}^T W \sim_{(W,D)} \rho_W(W)$, then $\theta(F,G,\hat{\beta})$ and $\theta(F,G,\beta^*)$ have the same asymptotic distribution.*

Proof. 1. The first claim is an application of the Neyman-Pearson Lemma, as pointed out by McIntosh and Pepe (2002) . Let an FPR value $\alpha \in (0, 1)$ be given. Viewing D as a parameter, the most powerful level α test of the null $D = 0$ versus the simple alternative $D = 1$ based on W rejects for large values of the likelihood ratio of (W, D), i.e., $f_{W|D=1}(W)/f_{W|D=0}(W)$. Therefore, the value of the ROC curve of the likelihood ratio at α, which is the power of the Neyman-Pearson test, is maximal. Since the ROC curve is the same for increasing functions of the likelihood, and the risk is the expit of the log likelihood, the same holds of the risk.

2. Though markers not related by an increasing function may have the same AUC, since the ROC curve of the risk is maximal, an index with the same AUC must have the same ROC curve. The latter does imply the index has the same conditional distributions as an increasing function of the risk.

□

Example 2 (Coefficient estimator: any non-zero estimate, covariate restrictions: A single covariate)**.** When there is a single covariate, $p = 1$, the β in (2), for $\beta \neq 0$, cancels and the requirement is simply that the risk be increasing in the sole covariate, i.e., that the covariate or its negation be a risk factor.

Example 3 (Parametric models where index is monotonically related to the risk)**.** The derivative will vanish in smooth parametric models under which the index is monotonically related to the risk function.

Example 3.1 (Coefficient estimator: binary response MLE, covariate restrictions: GLM link)**.** A prominent example where the index is an increasing function of the risk is the index model for a binary response:

$$P(D = 1 \mid W = w) = h(\beta^T w), \beta \in \mathbb{R}^p.$$

The function h is strictly increasing, such as a probit link, logistic link, identity, etc.

Example 3.2 (Coefficient estimator: linear discriminant analysis, covariate restrictions: multivariate Gaussian)**.** With $W \mid D = i \sim N(\mu_i, \Sigma), i = 1, 2$, the LDA estimate of β has probability limit $\beta^* = \Sigma^{-1}(\mu_1 - \mu_0)$. The likelihood ratio $f_{W|D=1}(w)/f_{W|D=0}(w)$ is an increasing function of $(\mu_1 - \mu_0)^T \Sigma^{-1} w = \beta^{*T} w$. That the derivative vanishes also follows by taking $\Sigma_0 = \Sigma_1$ in the upper bound given by Proposition 4.

Example 3.3 (Coefficient estimator: LDA, covariate restrictions: independent exponential family covariates with mean parameters, etc.)**.** Let component i of W, $i = 1, \ldots, p$, have conditional density given $D = j, j = 0, 1$, of the form $h_i(w) \exp(\theta_{ij} w - A_{ij})$. If the components are independent, the likelihood ratio then satisfies

$$\frac{f(w \mid D = 1)}{f(w \mid D = 0)} \sim (\theta_1 - \theta_0)^T w$$

With the usual LDA estimators, $\beta^* = \Sigma^{*-1} \Delta\mu^*$, where $\Delta\mu^* = A_1' - A_0'$ and Σ^* is diagonal with entries $\pi_0 A_{i0}'' + \pi_1 A_{i1}'', i = 1, \ldots, p, \pi_0 = 1 - \pi_1 = P(D = 0)$. If 1. the population variances are equal, i.e., $\pi_0 A''_{i0} + \pi_1 A''_{i1}$ doesn't depend on i, and 2. θ_i is the mean A'_i, then $\beta^{*T} w \sim (A'_1 - A'_0)^T w \sim \rho_W(w)$. This application of LDA is not justified under the usual LDA homoscedasticity assumption, as the parameter $\theta_j, j = 0, 1$, may affect the variance across classes. It turns out the coefficient estimation still does not affect the asymptotic distribution of the index AUC.

With Gaussian data as in Example 3.2 but unequal class variances, or heteroscedastic exponential family data as in 3.3 but non-independent covariates, the derivative of the AUC coefficient limit need not vanish, as shown by the example in Section 5.2.

5.1.2 Difference of AUCs

Next we consider application of the examples given in Section 5.1.1 to two AUCs computed from the same data, as when computing the difference.

Example 4. For a nonparametric estimator like MRC, Example 1, there is no further difficulty. Both estimation procedures have a vanishing derivative. This example is discussed in Heller et al. (2017).

Example 5. Likewise, there is no difficulty when one coefficient is estimated by a well-specified parametric estimator and the other by a nonparametric estimator, or e.g. the single covariate case where there is effectively no estimator. See, e.g., Fig. 1 in Demler et al. (2017) and the corresponding simulation.

Example 6 (Parametric models)**.** Next suppose both coefficient vectors being compared are modeled parametrically, and consider specifically nested binary response models 3.1. First, if neither the full model nor the reduced model is well spcified, there is no reason to expect the derivative to vanish by virtue of 3.1 and in general coefficient estimation must be accounted for. Second, if the reduced model is well-specified, then comparison with a superset of the covariates will generally lead to the null situation, i.e., a degenerate U-statistic, as discussed in Section 2. Finally, suppose that the full model is well-specified, e.g., when the full model contains a superset of the model covariates, and the reduced model a strict subset of the fuller set. In many cases correctness of the full model,

$$P(D = 1 \mid (W, W') = (w, w')) = h(\beta^T(w, w')), \text{ for some } \beta \in \mathbb{R}^{p+q} \tag{11}$$

implies the reduced model cannot be correct

$$P(D = 1 \mid W = w) = E(h(\beta^T(w, w')) \mid w) \neq h(\beta^T w) \text{ for any } \beta \in \mathbb{R}^p.$$

As a result the derivative term contributed by the reduced set AUC will be nonzero and must be accounted for. For the binary response model, the requirement is that the marginalization does not break the model, i.e., the inequality in 6 is an inequality for some β in the reduced model coefficient space. This condition is somewhat different from "collapsibility," the requirement that the coefficients belonging to the remaining covariates be the same after integrating out an independent set of covariates. Some examples where this condition holds are:

Example 6.1 (Coefficient estimator: probit regession, covariate restrictions: Gaussian)**.** When h in (11) is the standard normal CDF and the covariates are multivariate Gaussian, the marginalized model respects the probit link.

When h is the logistic function, and logistic regression is used to estimate the coefficients, the marginalized risk is not generally a logistic function of the index. Several authors have suggested, however, that the coefficient estimation may be ignored (Demler et al., 2011). In fact, it may be shown that if the covariates are Gaussian, then for any link function h, the adjustment term vanishes only when the coefficients under the marginalized model are the coefficients that would be obtained under a probit model. Therefore the adjustment term does not usually vanish when the coefficients are estimated under logistic regression. However, the difference may be negligible, so that in practice the coefficient estimation may be ignored. A heuristic reason to expect a small adjustment term is that the logistic risk is often close to the probit risk, where the adjustment does vanish.

Example 6.2 (Coefficent estimator: LDA, covariate restrictions: Gaussian)**.** When the full model is a well-specified Gaussian LDA model 3.2, the reduced model will be as well. This example is discussed in Demler et al. (2011) and considered further in Sections 5.2 and 6. LDA models are collapsible more generally, but the index may not be an increasing function of the risk/likelihood without the Gaussian assumption.

5.2 Coefficient estimation affects inference

Next we describe a situation where estimation of the coefficient must be accounted for when conducting inference on the difference of AUCs or even just a single AUC. The setting is a misspecification of the parametric model 3.2 where, but for the misspecification, coefficient estimation would not affect the asymptotic distribution. The nonparametric estimator provides the required adjustment missing from the basic Delong estimator, and may therefore be viewed as a robust estimator.

Suppose Gaussian linear discriminant analysis is applied to estimate the coefficient vector but the model is possibly misspecified in that the two classes may not have the same covariance. The model is:

$$\begin{aligned} W|D = d \sim F_d = N_p(\mu_d, \Sigma_d), \Sigma_d > 0, d \in \{0,1\}, \mu_1 \neq \mu_0 \\ P(D=1) = 1 - P(D=0) = \pi_1. \end{aligned} \tag{12}$$

The parameters are μ_d, Σ_d, and π_d, $d \in \{0,1\}$, for which sufficient statistics are given by the $2p(p+1)+1$ terms of $T = (\sum D_i, \sum D_i W_i {W_i}^T, \sum D_i W_i, \sum(1-D_i) W_i {W_i}^T, \sum(1-D_i)W_i)$. LDA bases class membership on the sign of $\hat{\beta}^T x$, where

$$\begin{aligned} \hat{\beta} &= \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0) \\ \hat{\mu}_1 - \hat{\mu}_0 &= N^{-1}\sum_i X_{1i} - M^{-1}\sum_i X_{0i} = T_3/T_1 \\ \hat{\Sigma} &= (M+N)^{-1}\left(\sum_i (X_{0i} - \hat{\mu}_0)(X_{0i} - \hat{\mu}_0)^T + \sum_i (X_{1i} - \hat{\mu}_1)(X_{1i} - \hat{\mu}_1)^T\right) \\ &= (M+N)^{-1}(T_2 - T_3 {T_3}^T/T_1 + T_4 - T_5 {T_5}^T/T_1). \end{aligned}$$

An intercept is usually computed when carrying out LDA but may be ignored here since the AUC is invariant to shifts. The LDA parameter estimates, under the unmet assumption of a common variance for the two classes, tend in probability to

$$\begin{aligned}\beta^* &= \Sigma^{*-1}(\mu_1 - \mu_0)\\ \Sigma^* &= \pi_0\Sigma_0 + \pi_1\Sigma_1\end{aligned}$$

Let $\Sigma = (\Sigma_0 + \Sigma_1)/2$. The index AUC and its partial derivative with respect to the coefficient vector, evaluated at the probability limit β^*, are

$$\begin{aligned}\theta(F,G,\beta^*) &= \Phi\left(\frac{\beta^{*T}\Sigma^*\beta^*}{\sqrt{2\beta^{*T}\Sigma\beta^*}}\right) > 1/2\\ \frac{\partial}{\partial\beta}\theta(F,G,\beta^*) &= (\pi_1 - \pi_0)\phi\left(\frac{\beta^{*T}\Sigma^*\beta^*}{\sqrt{2\beta^{*T}\Sigma\beta^*}}\right)\frac{\beta^{*T}}{(2\beta^{*T}\Sigma\beta^*)^{3/2}}((\beta^{*T}\Sigma_1\beta^*)\Sigma_0 - (\beta^{*T}\Sigma_0\beta^*)\Sigma_1).\end{aligned}$$

The derivative is 0 at β^* iff $(\beta^T\Sigma_0\beta)\Sigma_1\beta = (\beta^T\Sigma_1\beta)\Sigma_0\beta$ for $\beta = \beta^*$, equivalently, β^* is an eigenvector of $\Sigma_0^{-1}\Sigma_1$. For example, if the covariates are independent with a common variance, $\Sigma_d \propto I$, the derivative will be 0. As a second example, if $\Sigma_0 \propto \Sigma_1$, then $\Sigma_0^{-1}\Sigma_1 \propto I$ and again the derivative vanishes. The first example is already implied by the general exponential family result in Example 3.3 above but not the second as here the components are not necessarily independent, as required by that result. Even when the derivative is large, its effect may be mitigated. When $\Sigma \approx \Sigma^*$, the derivative term is approximately $O(|\Sigma|^{1/2})$, whereas the influence function is approximately $O(|\Sigma|^{-1/2})$, the root of the inverse Fisher information, so that the product, giving the entire adjustment term, is approximately $O(1)$.

The size of the adjustment term typically grows with the imbalance between the classes, as shown in Proposition 4. One therefore expects that under this scenario inference based on the Delong estimator will be faulty, as seen in the simulations in Section 6.

Proposition 4. *Let (W, D) follow (12), with full rank conditional covariance matrices $\Sigma_d, d = 0, 1$, and let $\psi_{\hat\beta}$ be the influence function of $\hat\beta$ based on a sample of size n. Then,*

1. *The standard deviation of the adjustment term is given by*

$$\begin{aligned}sd(\frac{\partial}{\partial\beta}\theta(F,G,\beta)\sqrt{n}\psi_{\hat\beta}) &= \left|(n\operatorname{Var}\psi_{\hat\beta})^{1/2}\frac{\partial}{\partial\beta}\theta(F,G,\beta)\right|\\ &= \phi\left(\frac{\beta^{*T}\Sigma^*\beta^*}{\sqrt{\beta^{*T}\Sigma\beta^*}}\right)\frac{|\pi_1 - \pi_0|}{(\beta^{*T}\Sigma\beta^*)^{3/2}}\left|\left(\Sigma^{*-1}\left(\Sigma_0/\pi_0 + \Sigma_1/\pi_1\right)\Sigma^{*-1}\right)^{1/2}\left((\beta^{*T}\Sigma_1\beta^*)\Sigma_0 - (\beta^{*T}\Sigma_0\beta^*)\Sigma_1\right)\beta^*\right|\end{aligned}$$

   Therefore the adjustment term vanishes when the classes are balanced, $\pi_0 = \pi_1 = 1/2$, or the model is homoscedastic, $\Sigma_0 = \Sigma_1$.

2. *The adjustment term vanishes iff the LDA parameter β^* is an eigenvector of $\Sigma_0^{-1}\Sigma_1$, equivalently, if $\mu_1 - \mu_0$ is an eigenvector of $\Sigma\Sigma^{*-1}$.*

3. *Consider the sub-model of* (12) *where the covariates are independent, i.e.,* Σ_0 *and* Σ_1 *are diagonal, and* $\pi_1 > 1/2$*. The parameters of this sub-model are* $\pi_1 > 1/2$*,* β*, and the diagonal entries of* Σ_0 *and* Σ_1*, denoted* Σ_{0ii} *and* Σ_{1ii}*,* $1 \le i \le p$*. Then*

$$\begin{aligned}&(n\,\mathrm{Var}\,\psi_{\hat{\beta}})^{1/2}\frac{\partial}{\partial\beta}\theta(F,G,\beta)\\ &=\phi\left(\frac{\beta^{*T}\Sigma^{*}\beta^{*}}{\sqrt{2\beta^{*T}\Sigma\beta^{*}}}\right)\frac{\pi_1-\pi_0}{(2\beta^{*T}\Sigma\beta^{*})^{3/2}}\left(\dots\sqrt{\frac{\Sigma_{0ii}}{\pi_0}+\frac{\Sigma_{1ii}}{\pi_1}}\,\frac{(\beta^{*T}\Sigma_1\beta^{*})\Sigma_{0ii}-(\beta^{*T}\Sigma_0\beta^{*})\Sigma_{1ii}}{\pi_0\Sigma_{0ii}+\pi_1\Sigma_{1ii}}\beta_i\dots\right)^t.\end{aligned}$$

Therefore if $\pi_0 \to 0$ *and for some* i $\Sigma_{1ii} \to 0$ *at the same time, with the remaining parameters held fixed, then the standard deviation of the adjustment term* $\to \infty$*.*

6 Simulation

We examine by simulation the coverage rate of the proposed and standard estimators of the index AUC and of the difference of AUCs. To do so we postulate nested models for the covariates on which the AUCs are based. Nesting models the practice of forming the AUC based on a set of covariates, removing one or more covariates, forming the AUC on the reduced set of covariates, and taking the difference of the AUCs. The analysis considers CIs for the full model AUC, reduced model AUC, and the difference. Besides the CI based on the proposed variance estimator, we also form CIs based on the standard Delong and bootstrap variance estimators. Also included is an "oracle" CI that uses the parametric form of the derivative, but estimates all other quantities. It is included primarily to illustrate the impact of nonparametrically estimating the derivative at the probability limit of the coefficient estimates.

6.1 Simulation settings

We consider two sets of nested models for the data. They are both sub-models of the heteroscedastic LDA model (12), and in fact the sub-model given in Proposition 4. These two sets of nested models are chosen to demonstrate that both poor coverage and poor power may result from failing to take into account the adjustment term (6).

1. Model 1. Reduced model: $\underset{p^{(r)}\times p^{(r)}}{\Sigma_0^{(r)}} = s^{(r)}I_{p^{(r)}}$, $s^{(r)} \in \mathbb{R}$, $\underset{p^{(r)}\times p^{(r)}}{\Sigma_1^{(r)}} = I_{p^{(r)}}$, and $\underset{p^{(r)}\times 1}{\beta^{(r)}} = b^{(r)}\mathbb{1}_{p^{(r)}}$, $b^{(r)} \in \mathbb{R}$. Full model: $\underset{p^{(f)}\times p^{(f)}}{\Sigma_0^{(f)}}$ is formed by appending $s^{(f)} > 0$ along the diagonal of $\Sigma_0^{(r)}$, $\underset{p^{(f)}\times p^{(f)}}{\Sigma_1^{(f)}} = I_{p^{(f)}}$, and $\underset{p^{(f)}\times 1}{\beta^{(f)}}$ is formed by appending $b^{(f)} \in \mathbb{R}$ to $\beta^{(r)}$. The parameters are $s^{(r)}, b^{(r)}, s^{(f)}, b^{(f)}$, all real-valued.

2. Model 2. Reduced model: $\underset{p^{(r)}\times p^{(r)}}{\Sigma_0^{(r)}}$ is diagonal with half the diagonal entries set to $\epsilon > 0$ and the other half $2s_0^{(r)} - \epsilon$, $s_0^{(r)} \in \mathbb{R}$, $\underset{p^{(r)}\times p^{(r)}}{\Sigma_1^{(r)}}$ is diagonal with half the diagonal entries

set to $\epsilon > 0$ and the other half $2s_1^{(r)} - \epsilon, s_1^{(r)} \in \mathbb{R}$, and $\underset{p^{(r)}\times 1}{\beta^{(r)}} = \mathbb{1}_{p^{(r)}}$. Full model: $\underset{p^{(f)}\times p^{(f)}}{\Sigma_0^{(f)}}$ and $\underset{p^{(f)}\times p^{(f)}}{\Sigma_1^{(f)}}$ are formed by appending $s_0^{(f)} > 0$ and $s_1^{(f)} > 0$, respectively, along the diagonals of $\Sigma_0^{(r)}$ and $\Sigma_1^{(r)}$, and $\underset{p^{(f)}\times 1}{\beta^{(f)}} = \mathbb{1}_{p^{(f)}}$. The parameters are $s_0^{(r)}, s_1^{(r)}, s_0^{(f)}, s_1^{(f)}, \epsilon$, all positive reals.

In the first model, the class covariance matrices $\Sigma_0^{(r)}, \Sigma_1^{(r)}, \Sigma_1^{(f)}$, are all proportional to the identity, and the difference of the AUCs between the full and reduced model is specified by solving for the components $b^{(r)}, b^{(f)}$ of β^*. In the second model, the full coefficient vector β^* is a constant vector, and the difference of the AUCs between the full and reduced model is specified by solving for the parameters $s_0^{(r)}, s_1^{(r)}, s_0^{(f)}, s_1^{(f)}$ determining the class covariance matrix. In both models, there is a nonzero adjustment term (6) in the difference of the AUCs when $\pi_1 \neq 1/2$. However, between these two models we can see two different ways in which this adjustment term may arise from the full and reduced models that underlie the difference in AUCs. Whereas in the second model both the full and reduced models require adjustment, in the first only the full model requires adjustment, the reduced model class covariance matrices being proportional to the identity matrix (Proposition 4).

In the results below, the full model and reduced models have $p^{(f)} = 7$ and $p^{(r)} = 6$ covariates, corresponding to an analysis of the impact of a single covariate on the AUC. The true AUC of the full model and reduced models are .8 and .77, so the true difference in AUCs is .03. The sample size is varied from 500 to 5,000. As discussed in Section 5.2, the magnitude of the adjustment required due to coefficient estimation is related to the class proportions in the LDA model, with no adjustment being required for balanced classes. We consider 3 levels of class imbalance, $\pi_1 = .5, .75, .9$. The sample size, difference in AUC, class imbalance, and other settings were informed by the data analysis in Demler et al. (2017) and may be easily changed in the supplied software.

6.2 Simulation results

Results are presented in Figure 1 and 1. When there is no class imbalance, $\pi_1 = 1/2$, the estimators approximate the nominal rate, with the exception of the bootstrap under Model 2. The CI lengths of the proposed estimator are nearly identical to the Delong estimator, suggesting there is little loss of efficiency in estimating the adjustment to be 0.

When a class imbalance is present, as expected from Proposition 4, the Delong estimator, which does not take into account the coefficient estimation, does not approximate the nominal rate, with performance deteriorating with the magnitude of the imbalance. The bootstrap does not fare well either. Under Model 1, the Delong CI falls far below the nominal rate, whereas the bootstrap CI is underpowered. Under Model 2, the situation is reversed.

The results for Model 2 suggest that the FPR under the bootstrap, though far below the nominal rate, appears to improve with sample size. It does appear from further simulation, not presented here, that the bootstrap does eventually approximate the nominal rate. In these simulations the bootstrap's performance given sample sizes on the order of $n = 100k$ is become comparable to the proposed estimator's at the sample sizes presented in Figure 1, corresponding to a relative efficiency on the order of 1/20. However, unlike the Delong estimator, the bootstrap estimator does appear to be consistent, though its inefficiency is prohibitive for many settings.

The proposed estimator approximates the nominal rate. There does not appear to be much difference between the oracle and proposed estimators, suggesting nonparametric approximation of the gradient at β^* is not a problem at these sample sizes.

7 Data Analysis

We re-analyze the Framingham Heart Study set analyzed in Demler et al. (2011, 2017). The FHS followed 8,223 individuals free of cardiovascular disease at baseline. The outcome of interest was the presence of coronary heart disease in a 12-year follow-up, which 7.5% of the participants developed. The FHS data forms the basis of the Framingham Risk Score for 10-year risk of cardiovascular disease. More information on the study and risk score may be found in D'Agostino Sr et al. (2008).

We compare how well several indices based on logistic regression discriminate between individuals who did and did not develop CHD. Specifically, we form CIs for $\Delta\theta = \theta - \theta'$ where the AUCs θ is an index AUC where the index is based on a full set of covariates and θ' is an index AUC obtained after omitting a covariate from the full set. Let $D_i = 1$ or $= 0$ as subject i did or did not develop CHD, let w_i denote the full vector of subject i's predictors, and let $x_1, \ldots, x_m$, and $y_1, \ldots, y_n$, enumerate these predictors for control and case subjects. The full set of covariates w consists of a smoking indicator, age, systolic and diastolic blood pressure, a diabetes indicator, and total and high-density lipoprotein cholesterol. The coefficient vector $\hat{\beta}$ is obtained by fitting the logistic regression $P(D = 1) = \sigma(\alpha + \beta^T w)$, where σ is the logistic function. The reduced set of vectors w' is obtained by omitting one of the predictors in w, and the coefficient vector $\hat{\gamma}$ is obtained by fitting the logistic regression $P(D = 1) = \sigma(\alpha' + \gamma^T w')$.

CIs were formed using the variance estimate given by Corollary 2 as well as the unadjusted Delong estimator. The CI given by Corollary 2 assumes $\beta^* \neq \gamma^*$. The reduced sets of covariates are therefore formed by omitting only significant covariates from the full set w, where the significance is determined by logistic regression on the full set. All covariates in w except systolic and total blood pressure are significant at the 5% level. This approach is consistent with the guidance given in Pepe et al. (2013) to assess significance first and only then consider CIs for $\Delta\hat{\theta}$.

The resulting CIs are plotted in Figure 2. The CIs formed with the adusted estimator are similar to the unadjusted estimator throughout. This result corroborates the argument given in Demler et al. (2011, 2017) that no adjustment is necessary when the index AUC coefficients are obtained using logistic regression: No adjustment is necessary when $\hat{\beta}, \hat{\gamma}$ are obtained by LDA (see Example 3.2), and as the coefficients in a well-specified logistic regression are known to have the same probability limit as LDA coefficients, the same holds when $\hat{\beta}, \hat{\gamma}$ are obtained by logistic regression, at least asymptotically. This argument does not seem in and of itself to justify the conclusion that no adjustment is necessary. While well-specified logistic and LDA models coefficients will converge to the same limit, it is typically not the case that a logistic model and reduced logistic model obtained after omitting covariates are simultaneously consistent. Unlike LDA, the logistic model is not collapsible (see Example 6).

As a consequence, one of the index AUCs that show up in the differnce, either based on $\hat{\beta}$ or $\hat{\gamma}$, will require adjustment. Nevertheless, the conclusion that the adjustment is negligible does appear to hold. A reason suggested in Example 6 is that probit regression is collapsible and therefore requires no adustment if well-specified, and the difference between logistic and probit coefficients is negligible.

8 Discussion

We have described a nonparametric method of estimating the index AUC when the index coefficients are estimated from the same data on which the AUC estimate is based. The method applies directly to testing for the difference of index AUCs with estimated coefficients, when the two AUCs are in the limit distinct. The method described above applies not only to testing indexes based on nested data sets, perhaps the most common situation, but more generally to a comparison of any correlated AUCs with index coefficients estimated from the data, e.g., LDA versus logistic. The method is easily extended to other differentiable functions of the data, not just an index. For example, in the heteroscedastic LDA example considered in Section 5.2, a common solution would be to use quadratic discriminant analysis, though the marker in this case would be a quadratic function of the covariates. An important limitation of the method described here is the requirement that the coefficient estimation procedure have an influence function, excluding many modern classification techniques.

While we have described inference that accounts for the adjustment ignored by the standard error, we have also described many situations in which no adjustment is needed. Even in the heteroscedastic LDA model, where the adjustment term can be arbitrarily large, the unadjusted estimator performs reasonably when the class imbalance across cases and controls is not very extreme. These results justify the use of the standard Delong estimator in spite of the unmet assumption of IID data, but they typically require parametric assumptions. The proposed estimator may therefore be viewed as a robust estimator, enabling inference despite departures from the assumed model. From the presented simulation results, there is not much loss of efficiency in using the robust estimator.

References

DeLong, E. R., D. M. DeLong, and D. L. Clarke-Pearson (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 837–845.

Demler, O. V., M. J. Pencina, N. R. Cook, and R. B. D'Agostino Sr (2017). Asymptotic distribution of δauc, nris, and idi based on theory of u-statistics. *Statistics in Medicine 36*(21), 3334–3360.

Demler, O. V., M. J. Pencina, and R. B. D'Agostino Sr (2011). Equivalence of improvement in area under roc curve and linear discriminant analysis coefficient under assumption of normality. *Statistics in medicine 30*(12), 1410–1418.

Demler, O. V., M. J. Pencina, and R. B. D'Agostino Sr (2012). Misuse of delong test to compare aucs for nested models. *Statistics in medicine 31*(23), 2577–2587.

D'Agostino Sr, R. B., R. S. Vasan, M. J. Pencina, P. A. Wolf, M. Cobain, J. M. Massaro, and W. B. Kannel (2008). General cardiovascular risk profile for use in primary care: the framingham heart study. *Circulation 117*(6), 743–753.

Heller, G., V. E. Seshan, C. S. Moskowitz, and M. Gönen (2017). Inference for the difference in the area under the roc curve derived from nested binary regression models. *Biostatistics 18*(2), 260–274.

Hoeffiding, W. (1948). A class of statistics with asymptotically normal distributions. *Annals of Mathematical Statistics 19*(3), 293–325.

Lee, C. Y. (2021). Nested logistic regression models and δauc applications: Change-point analysis. *Statistical Methods in Medical Research 30*(7), 1654–1666.

McIntosh, M. W. and M. S. Pepe (2002). Combining several screening tests: optimality of the risk score. *Biometrics 58*(3), 657–664.

Michael, H. et al. (2023). Inference on the difference of index AUCs under the null. Forthcoming; available at https://www.umass.edu/mathematics-statistics/directory/faculty/haben-michael.

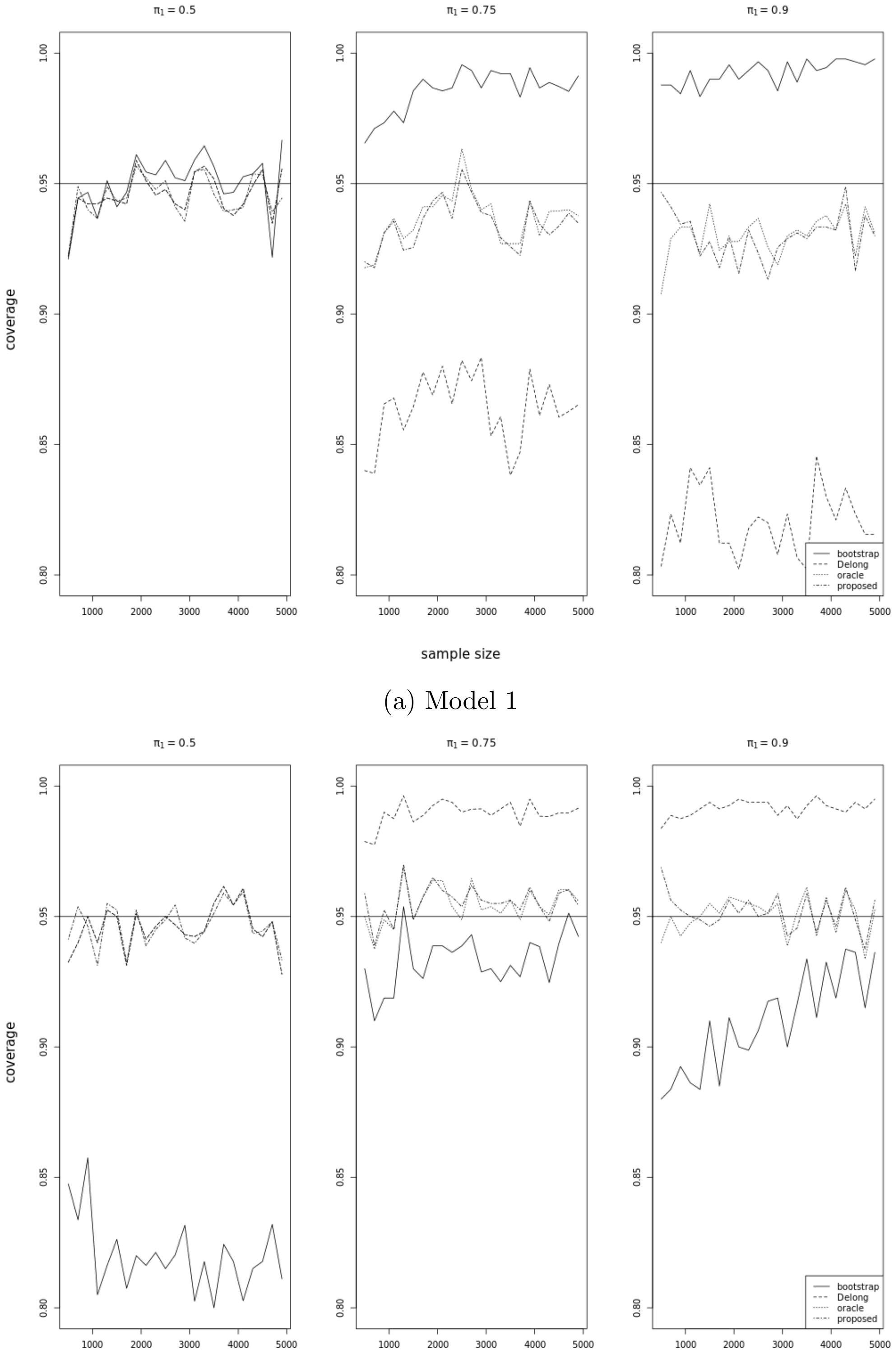
Pepe, M. S., K. F. Kerr, G. Longton, and Z. Wang (2013). Testing for improvement in prediction model performance. *Statistics in medicine 32*(9), 1467–1482.

Pollard, D. (1984). *Convergence of Stochastic Processes.*

Seshan, V. E., M. Gönen, and C. B. Begg (2013). Comparing ROC curves derived from regression models. *Statistics in medicine 32*(9), 1483–1493.

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

Tzoulaki, I., G. Liberopoulos, and J. P. Ioannidis (2009). Assessment of claims of improved prediction beyond the framingham risk score. *JAMA 302*(21), 2345–2352.



(a) Model 1

(b) Model 2

Figure 1: Simulation comparing the coverage of nominal 95% CIs for the difference in AUCs. In (a) the Delong CI has poor FPR and the bootstrap CI has poor power. The situation is reversed in (b).

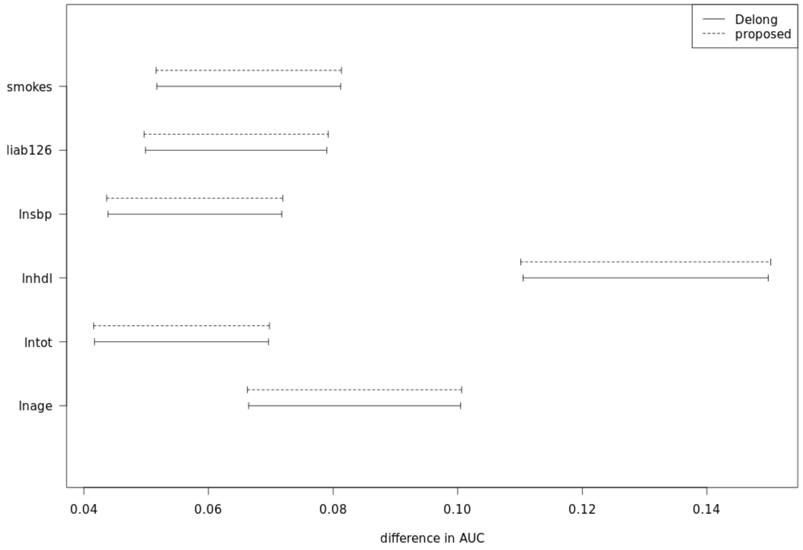


Figure 2: Confidence intervals for the difference in index AUCs estimated using the FHS data. One index AUC is computed using a full set of 7 covariates, and the other is obtained by omitting one of 6 variables. The adjusted and unadjusted estimators are nearly the same, as conjectured by Demler et al. (2017).

	n	500	2500	4500
imbalance	estimator			
0.1	bootstrap	0.988 (0.050)	0.997 (0.026)	0.997 (0.021)
	Delong	0.803 (0.027)	0.822 (0.012)	0.823 (0.009)
	oracle	0.908 (0.036)	0.937 (0.016)	0.922 (0.012)
	proposed	0.947 (0.044)	0.923 (0.017)	0.917 (0.012)
0.25	bootstrap	0.966 (0.042)	0.996 (0.022)	0.987 (0.017)
	Delong	0.840 (0.028)	0.882 (0.013)	0.860 (0.009)
	oracle	0.918 (0.035)	0.963 (0.015)	0.940 (0.012)
	proposed	0.920 (0.036)	0.956 (0.015)	0.934 (0.011)
0.5	bootstrap	0.921 (0.043)	0.959 (0.020)	0.958 (0.015)
	Delong	0.922 (0.043)	0.948 (0.019)	0.956 (0.015)
	oracle	0.922 (0.043)	0.948 (0.019)	0.956 (0.015)
	proposed	0.922 (0.044)	0.951 (0.019)	0.953 (0.014)

(a) Model 1

	n	500	2500	4500
imbalance	estimator			
0.1	bootstrap	0.880 (0.056)	0.906 (0.028)	0.936 (0.023)
	Delong	0.984 (0.097)	0.994 (0.045)	0.994 (0.034)
	oracle	0.940 (0.073)	0.954 (0.033)	0.953 (0.025)
	proposed	0.969 (0.080)	0.950 (0.033)	0.949 (0.025)
0.25	bootstrap	0.930 (0.046)	0.939 (0.022)	0.940 (0.017)
	Delong	0.979 (0.067)	0.990 (0.030)	0.990 (0.023)
	oracle	0.950 (0.053)	0.949 (0.024)	0.960 (0.018)
	proposed	0.959 (0.056)	0.954 (0.024)	0.959 (0.018)
0.5	bootstrap	0.848 (0.032)	0.815 (0.014)	0.818 (0.010)
	Delong	0.932 (0.043)	0.950 (0.020)	0.942 (0.015)
	oracle	0.932 (0.043)	0.950 (0.020)	0.942 (0.015)
	proposed	0.941 (0.046)	0.949 (0.020)	0.944 (0.015)

(b) Model 2

Table 1: Simulation comparing the coverage of nominal 95% CIs for the difference in AUCs. In (a) the Delong CI has poor FPR and the bootstrap CI has poor power. The situation is reversed in (b).