Nonparametric Estimation of the AUC of an Index with Estimated Parameters

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.

1 Introduction

The AUC is a standard measure of how effectively a marker discriminates between two classes. The difference in AUCs, written $\Delta\theta$, is a standard comparison of 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 additionatal factors $y \backslash x$. The difference in AUCs has been described by experts as one of the most widely used measures of discrimination (Demler et al., 2017). should be "difference" in discrimination I think?

A related way of measuring the impact of additional covariates is to directly compare the coefficients β and γ under a model. 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 ΔAUC (Seshan et al., 2013) johnny cite . While the source of the contradiction was soon identified, remedies have been slower to arrive.

We present here a partial remedy, a method to nonparametrically estimate $\Delta\theta$ under the assumption that $\Delta\theta \neq 0$. is this the right characterization of the null? useful cites: maybe cite seshan counts of clinical papers, sas proc. ((seshan 2013: In the first four months of 2011 alone, we easily identified seven articles in clinical journals that used the AUC test to compare nested logistic regression models [11, 12, 13, 14, 15, 16, 17]))

2 Background/setting

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

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

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 $(W_1, D_1), \ldots, (W_{m+n}, D_{m+n})$, be an IID sample under (1), with the class variables

$$X_{01}, \ldots, X_{0m} \overset{IID}{\sim} F, X_{11}, \ldots, X_{1n} \overset{IID}{\sim} G, m = \sum\{D = 1\}, n = \sum\{D = 0\}$$

Vectors $\hat{\beta} \in \mathbb{R}$ and $\hat{\gamma} \in \mathbb{R}$ are obtained based on the sample by some means. 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 maker associated with one class is less than a stochastically independent marker associated with the other class. A non-parametric estimator is the sample proportion of markers in one class less than the markers in the other. In the case that the markers are indexes with esimated coefficient $\hat{\beta}$, the estimator is

$$\frac{1}{mn} \sum_{i,j} \{\hat{\beta}^T X_{0i} < \hat{\beta}^T X_{1j}\}. \tag{2}$$

The difference of index AUCs is estimated nonparametrically by

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

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 cite applied papers , 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 $(x, y) \mapsto \{\beta^T x < \beta^T y\} - \{\gamma^T x < \gamma^T y\}$, 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 obesrvations. 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 method such as logistic regression remains "intractable" (Lee, 2021).

Instead, the literature has proposed the use of more covenient testing problems equivalent in certain settings to the 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 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, may as well test risk functions.

As we don't offer any such improvements over testing the risk, we only consider the alternative case $\Delta\theta \neq 0$ in the remainder.

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 Method

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.

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

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

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,

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

Where $\delta F=\hat{F}-F$, etc.

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

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

The third and final term in (7) is an average of mn uncorrelated terms and therefore $o(n^{-1/2})$.

For fixed $\beta^*+\delta\beta$, the first two terms 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 $\beta\mapsto\theta(\delta F,G,\beta),\beta\mapsto\theta(F,\delta G,\beta)$ (Pollard, 1984).

Therefore,

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

This IID representation is known as the Hoeffding decomposition of a U-statistic, and is the same as the first von Mises derivative. The CLT may be applied to get the asymptotic distribution of $\Delta\hat{\theta}$ in situations that the term (6) is negligible, e.g., if $\hat{\beta}=\beta$ were not estimated (see Section 5 for additional scenarios when (6) is negligible)). The approach of DeLong et al. (1988) in such situations is to estimate F,G above using the empirical CDFs $\hat{F},\hat{G}$, giving rise to the standard Delong statistic for inference on the $\Delta\theta$,

$$-\frac{1}{m}\sum_{i=1}^{m}(1-\hat{G}(\beta^TX_{0i})-\theta(\hat{F},\hat{G},\beta))+\frac{1}{n}\sum_{i=1}^{n}(\hat{F}(\beta^TX_{1i})-\theta(\hat{F},\hat{G},\beta)).$$

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 the $\hat{\beta}$. 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,

$$\begin{aligned}
&\theta(\hat{F}, \hat{G}, \hat{\beta}) - \theta(F, G, \beta) \\
&= \frac{1}{m+n}\sum_{i=1}^{m+n}\left(-\frac{\{D_i = 0\}}{P(D=0)}(1 - G(\beta^{*T}W_i) - \theta(F, G, \beta^*)) + \frac{\{D_i = 1\}}{P(D=1)}(F(\beta^{*T}W_i) - \theta(F, G, \beta^*))\right) \\
&+ \frac{\partial}{\partial\beta}\theta(F, G, \beta)^T\sum_{i=1}^{m+n}\psi_{\hat{\beta}}(W_i) + o_P((m+n)^{-1/2}) \qquad (9)
\end{aligned}$$

maybe combine into one sum

Proposition 1. *Given a sample $(W_1, D_1), \ldots, (W_{m+n}, D_{m+n})$ as (1), and estimator $\hat{\beta}$ based on the sample. Assumptions:*

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

Assertion: $(m+n)^{-1/2}(\theta(\hat{F}, \hat{G}, \hat{\beta}) - \theta(F, G, \beta^*))$ is asymptotically normal with mean zero and variance given by the variance of a term in (9). This variance may be consistently estimated as $\sqrt{m+n}$ times the sample variance of the terms in (9).

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 a sample $(W_1, D_1), \ldots, (W_{m+n}, D_{m+n})$ as (1), and estimators $\hat{\beta}, \hat{\gamma}$ based on the sample. Assumptions:*

1. *Assumptions of Proposition 1 apply to $\hat{\beta}, \hat{\gamma}$ both*
2. *$\beta^* \neq \gamma^*$*

Then $(m+n)^{-1}(\Delta\hat{\theta} - \Delta\theta)$ is asymptotically normal with mean zero and variance given by a term in the difference of IID means (9). This variance may be consistently estimated as $\sqrt{m+n}$ times the sample variance of the difference of terms as in (9)

It isn't required that $\hat{\beta}$ be estimated by a correctly specified model, only that it has some probability limit at the parametric rate. Though the procedure for obtaining the estimate $\hat{\beta}$ and the associated influence function ψ often involve some parametric assumptions, we still term the procedure described here as "non-parametric" 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.

What goes wrong under the null? If $\beta^* = \gamma^*$ then the main terms in the Hoeffding-type decomposition (8) 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. check. also check against demler 2017 iff conditions for degeneracy.

restructure–the above two paras are more like remarks. first 1 maybe goes to an estimation section.

4 Estimation

Using Proposition 1 for non-parametric inference will usually require certain parameters to be estimated. As mentioned in Section 3, even if the coefficient beta were not estimated, the terms in the estimator corresponding to the Delong statistic would still require estimating the CDFs. In the adjustment term, the influence function and derivative term will often require estimated parameters. Substituting consistent etimated parameters into the IID average is usually asymptotically negligible as long as the dependence is continuous, though of course the efficiency of the convergence may be affected. Non-parametric 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.

5 Examples

add linear regression to collapsibility?

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 estimtion of the coefficient may be ignored. In this case the usual Mann-Whitney U-statistic, in the case of $\hat{\theta}$, 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. When the adjustment term does not vanish, 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, $\partial_3\theta(F, G, \beta^*) = 0$ 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, $\partial_3\theta(F, G, \beta^*) = \partial_3\theta(F, G, \gamma^*) = 0$.

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 non-parametric. By construction the empirical AUC is stationary at the coefficient estimates, and under regularity conditions the AUC $\theta(F, G, \beta)$ is stationary at the probability limit β^*, as well.

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 $w \mapsto f_{W|D=1}(w)/f_{W|D=0}(w)$, cond desnities not defined 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 ${\beta^*}^T W \sim_{(W,D)} \rho_W(W)$, slight abuse of $\sim$ notation here 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 Pepe et al. (2013) check if she did it first. swets. . 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: LDA, 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 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 $D = j, j = 1, 2$, 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, $\pi_0 A_{i0}'' + \pi_1 A_{i1}''$ 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 homoskedasticity assumption, as the parameter $\theta_j, j = 1, 2$, 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 heteroskedastic 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 non-parametric estimator like MRC, Example 1, there is no further difficulty. Each estimation procedure leads to 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 parameric estimator and the other by a non-parametric 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. give citations to simulations/data anlayses In many cases correctness of the full model,

$$P(D = 1 \mid (W, W') = (w, w')) = h(\beta^T(w, w')), \text{ for some } \beta \in \mathbb{R}^{p+q} \tag{10}$$

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 (10) 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) give citations. is this the right demler paper? . In fact, as shown in forthcoming work or add to appendix? , for whatever link h, if the covariates are gaussian, then the adjustment

term vanishes, and coefficient estimation may be ignored, only when the covariates under the marginalized model are the covariates that would be obtained under a probit model. Though the derivative may not vanish, however, it 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 are 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) give their mahalanobis distance characterization. cite other papers that discuss this model 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 requires 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:

$$\begin{aligned} W|D = d \sim F_i = N_p(\mu_i, \Sigma_i), d = 1, 2 \\ P(D = 1) = 1 - P(D = 0) = \pi_1 \end{aligned} \tag{11}$$

The LDA parameter estimation procedure is to base class membership on the sign of $\hat{\beta}^T x$, where

$$\begin{aligned} \hat{\beta} &= \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0) \\ \hat{\mu}_1 - \hat{\mu}_0 &= n^{-1}\sum_i X_{1i} - m^{-1}\sum_i X_{0i} \\ \hat{\Sigma} &= m/(m+n)\sum_i (X_{0i} - \hat{\mu}_0)(X_{0i} - \hat{\mu}_0)^T + n/(m+n)\sum_i (X_{1i} - \hat{\mu}_1)(X_{1i} - \hat{\mu}_1)^T \end{aligned}$$

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 $\Sigma = (\Sigma_0 + \Sigma_1)/2, \Sigma_\pi = \pi_0\Sigma_0 + \pi_1\Sigma_1$. The index AUC and its partial derivative with respect to the coefficient are

$$\theta(F, G, \beta) = \Phi\left(\frac{\beta^T\Sigma_\pi\beta}{\sqrt{2\beta^T\Sigma\beta}}\right)$$
$$\frac{\partial}{\partial\beta}\theta(F, G, \beta) = \phi\left(\frac{\beta^T\Sigma_\pi\beta}{\sqrt{2\beta^T\Sigma\beta}}\right)\frac{\beta^T}{\sqrt{2}(\beta^T\Sigma\beta)^{3/2}}((\beta^T\Sigma_0\beta)\Sigma_1 - (\beta^T\Sigma_1\beta)\Sigma_0).$$

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 the observations are not independent. 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 $O(1)$.

As the imbalance between the classes increases, the size $|(\Sigma^*)^{-1}\theta'(F, G, \beta^*)|$ of the adjustment term may grow, as show in Proposition 4. One therefore expects that under this scenario inference based on the Delong estimator will be faulty, as verified by simulation in Section 6.

Proposition 4. *Let (W, D) follow (11), with full rank conditional covariance matrices $\Sigma_d, d = 0, 1$. Then,*

1. *The adjustment term satisfies*

$$sd(\frac{\partial}{\partial\beta}\theta(F, G, \beta)\psi_{\hat\beta}(W_i)) = |\Sigma_\pi^{-1/2}\frac{\partial}{\partial\beta}\theta(F, G, \beta)|$$
$$\leq |\pi_1 - \pi_0|/(2\sqrt{\pi})\frac{\lambda_1(\Sigma_0)|\Sigma_1 - \Sigma_0|}{(\lambda_n(\Sigma_0) + \lambda_n(\Sigma_1))^{3/2}(\pi_0\lambda_n(\Sigma_0) + \pi_1\lambda_n(\Sigma_1))^{1/2}}$$

   where $\lambda_j(\Sigma_i)$ is the jth largest eigenvalue of a $\Sigma_i, i = 1, 2$.

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

3. *Suppose p is even, $1/2 < \pi_0 < 1$, Σ_0 is a diagonal matrix with half the entries on the diagonal equal to $\epsilon \in (0, 2)$ and the other half $2-\epsilon$, Σ_1 is the identity, and $\beta = \sqrt{2/p}\mathbb{1}$. As $\pi_0 \to 1$ and $\epsilon \to 0$ simultaneously, just use ones vector below? pm should be outside square root?*

$$|\Sigma_\pi^{-1/2}\frac{\partial}{\partial\beta}\theta(F, G, \beta)| = |\frac{(\pi_1 - \pi_0)(1 - \epsilon)}{4\sqrt{p\pi e}}(1 + (\epsilon - 1)\pi_0)^{-1/2}\mathbb{1}| \to \infty.$$

6 Simulation

maybe give nnot in terms of adjustment size but class imbalance We examine by simulation the coverage rate of the proposed estimator of $\Delta\theta$ under the misspecified LDA model (11). We consider nested models, with 8 covariates in the full model and 6 in the reduced, varying the sample size from 1000 to 3000. These choices were informed by the data analysis in Demler et al. (2017).We consider 3 levels of magnitude for the adjustment term, $|...|_\infty = .05, .1, .12$. For comparison, we also consider coverage of the unadjusted delong estimator, and an oracle estimator, obtained using the derivative obtained in reliance on the paramtric family of the data (11).

Results are presented in Figure 1. As expected from Proposition 4, the Delong estimator, which does not take into account the coefficient estimation, falls far below the nominal coverage rate, with performance deteriorating with the size of the adjustment term. The proposed estimator approximates the nominal rate.

((maybe include L1 distance to Z?))

7 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 main results also apply to discrete covariates, though they were not considered among the examples here. The method is easily extended to other differentiable functions of the data, not just an index. For example, in the heteroskedastic LDA example considered above, 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 that the coefficient estimation procedure have an influence function, excluding many modern classification techniques.

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.

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.

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.

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*. David Pollard.

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

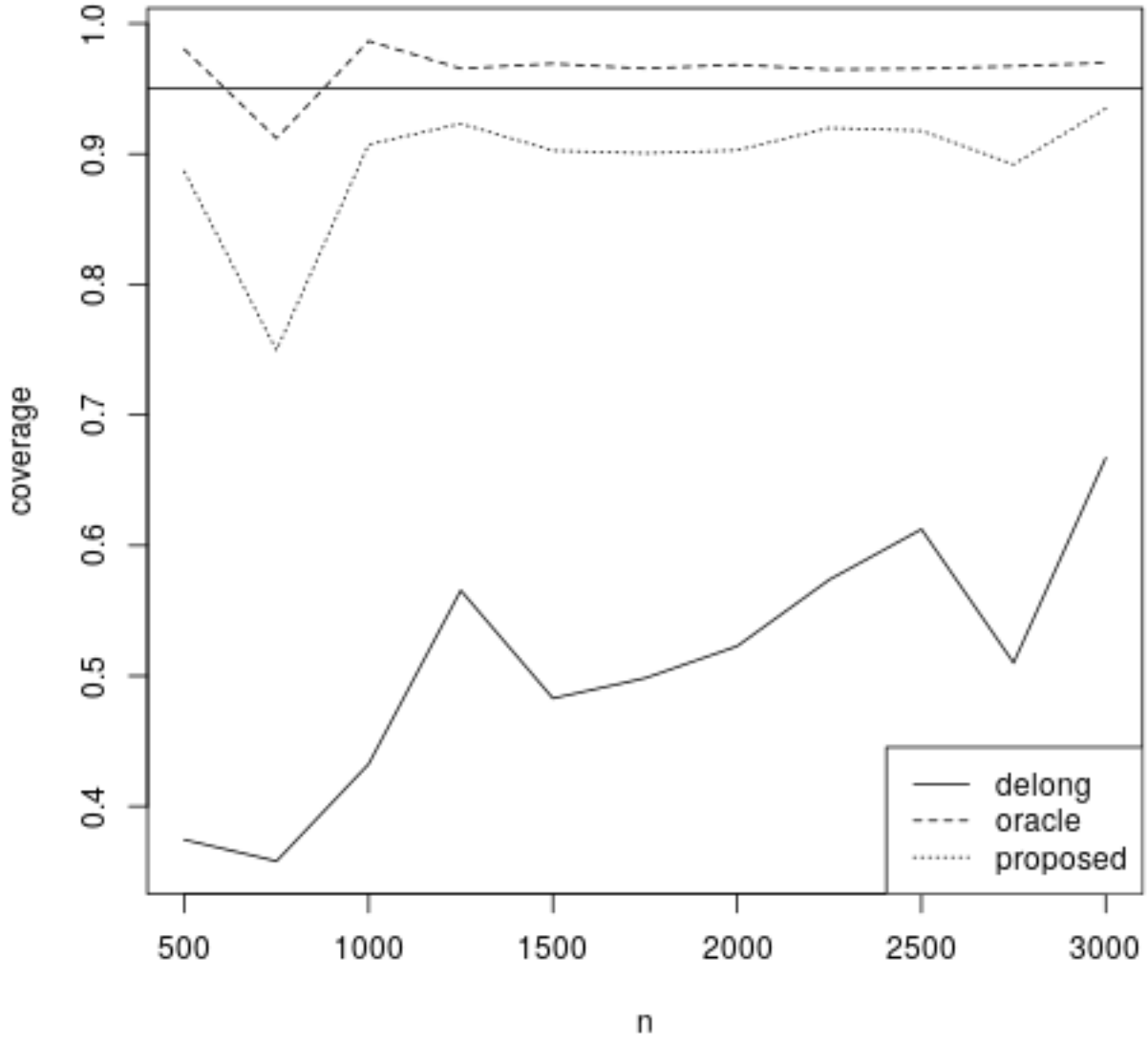

Figure 1: Coverage simulation