Inference on the Difference of AUCs Based on Fitted Values under the Null

Haben Michael
University of Massachusetts

Abstract: The difference of AUCs is a widely used measure of the improvement in class discrimination when comparing predictors. The predictors often take the form of indexes, the linear fitted values $\hat{\beta}^T w$ from some estimation procedure. Since the estimation procedure is often carried out using the same data as used to estimate the difference of AUCs, standard results on the distribution of the difference of AUCs, assuming independent observations, generally do not apply. Recent work has developed non-parametric inference procedures under the assumption that the true difference is nonzero, which is useful, e.g., for forming confidence intervals. The distribution under the assumption that the true difference is zero is of central importance for testing. However, the analysis is more complicated as the asymptotic distribution of the test statistic is generally non-normal, and only special cases have appeared in the literature. The asymptotic distribution is presented here under general conditions and parametric and non-parametric estimation are described. The previously published special cases are re-derived. In so doing we resolve a long-standing problem in the biomedical literature that has as recently as 2021 been described as "intractable."

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, and comparison of markers often takes the form of comparing the AUCs of two sets of patient characteristics x and 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).

Despite its widespread adoption, inferences on the difference in AUCs have been observed to be faulty (Tzoulaki et al., 2009; Seshan et al., 2013; Demler et al., 2012, 2017; Heller et al., 2017; Lee, 2021). The reason lies in the typical way in which the difference is estimated, where the same data is used both to obtain the coefficient vectors $\hat{\beta}$ as to estimate the difference of the AUCs of the fitted values $\hat{\beta}^T x$. The asymptotic null distribution of the test statistic formed in this way is in general non-normal. In certain special cases few solutions have appeared, but these do not cover important cases such as when $\hat{\beta}$ is estimated by logistic regression. We describe the asymptotic null distribution under general conditions and propose procedures for inference.

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 the asymptotic null distribution of the difference of AUCs based on fitted values. In Section 4 we describe parametric and non-parametric estimation of this distribution. Along the way we re-derive the special case of Heller et al. (2017). In Section 5 we examine the finite-sample performance of the proposed estimation procedures. Here we also re-analyze the special case given in Demler et al. (2011). We conclude and suggest extensions and directions for future work in Section 6. Software implementing the proposed inference procedure and the routines used in the simulation section are publicly available at the corresponding 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 $X \in \mathcal{X}, X \sim F, Y \in \mathcal{Y}, Y \sim G$ the RVs, state spaces, 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 control and case variables

$$X_1, \ldots, X_M \overset{IID}{\sim} F, Y_1, \ldots, Y_N \overset{IID}{\sim} G, M = \sum\{D = 0\}, N = \sum\{D = 1\}. \tag{2}$$

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 fixed $\sqrt{M + N}$-rate 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 a control marker is less than a stochastically independent case marker, with ties weighted by half. A nonparametric estimator of the AUC is the sample proportion of control markers less than case markers, with ties weighted by half. In the case that the markers are indexes with estimated coefficient $\hat{\beta}$, the estimator takes the form

$$\hat{\theta} = \frac{1}{MN} \sum_{i,j} \psi(\hat{\beta}^T X_i, \hat{\beta}^T Y_j),$$

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

$$\Delta = E\psi(\beta^{*T} X, \beta^{*T} Y) - E\psi(\gamma^{*T} X, \gamma^{*T} Y)$$

is estimated nonparametrically by

$$\hat{\Delta} = \frac{1}{MN} \sum_{i,j} \psi(\hat{\beta}^T X_i, \hat{\beta}^T Y_j) - \frac{1}{MN} \sum_{i,j} \psi(\hat{\gamma}^T X_i, \hat{\gamma}^T Y_j).$$

The asymptotic distribution of $\hat{\Delta}$ is sought for inference. The proper normalization of $\hat{\Delta}$ depends on the probability limits β^* and γ^*. When $\beta^* \neq \gamma^*$ a $\sqrt{M+N}$ normalization is commonly appropriate, leading to an asymptotically normal distribution. This situation is discussed in Doyle-Connolly and Michael (2023). The distribution when $\beta = \gamma$ is considered in this paper. In this situation $\hat{\Delta}$ must often be normalized by $M+N$ to obtain a proper asymptotic distribution, which is then a combination of products of normals. Whereas the distribution of $\hat{\Delta}$ when $\beta \neq \gamma$ is useful for forming confidence intervals for Δ, the distribution under $\beta = \gamma$ is useful for testing the null hypothesis of no difference between the AUCs of the two markers,

$$H_0 : \Delta = 0. \tag{3}$$

A number of early papers (Demler et al., 2011, 2012; Seshan et al., 2013) observed that the standard test (DeLong et al., 1988) for the difference of index AUCs did not return anticipated results, e.g., a non-significant p-value when one marker was otherwise known to be significantly more informative than the other. Researchers soon identified the use of estimated coefficients as the culprit and, using synthetic data, observed that the limiting distribution appeared non-normal. Several parametric results followed, making assumptions about the distribution of the covariates or the coefficient estimation procedure, or both. Demler et al. (2011) observed that when the covariates are normal and the coefficients are estimated by linear discriminant analysis (see Section 4.1.1), the null $\Delta = 0$ may be tested with an F-test. Pepe et al. (2013) gave the far-reaching result that testing for a difference of AUCs of risk functions is the same as testing for a difference in the risk functions. Heller et al. (2017) derive the asymptotic distribution of $\hat{\Delta}$ when the coefficients are estimated by the maximum rank correlation (see Section 4.1.2). Remarking that the asymptotic null distribution is intractable in common cases, Lee (2021) proposes a resampling approach when the data satisfy a change-point assumption.

Finally, Sherman (1993) considers the generalized regression model proposed by Han (1987), with the object of obtaining asymptotic normality of an estimator for the model parameters. This estimator is the maximum rank correlation, which is based on an index AUC. Although the paper does not seem to be a part of the more recent literature on the difference in AUCs, many of the methods are similar to those used here.

3 Theory

Given $\beta, \gamma \in \mathbb{R}^p$ and distributions F and G on $\mathbb{R}^p$, let

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

and

$$\Delta(\beta, \gamma, F, G) = \int \psi(\beta^T x, \beta^T y) dF(x) dG(y) - \int \psi(\gamma^T x, \gamma^T y) dF(x) dG(y)$$

denote respectively the index AUC and the difference of index AUCs written as statistical functionals. Let $\hat{F}, \hat{G}$ denote the empirical CDFs of the control and case observations. With this notation $\hat{\Delta} = \Delta(\hat{\beta}, \hat{\gamma}, \hat{F}, \hat{G})$ and the null hypothesis (3) may be written $\Delta(\beta^*, \gamma^*, F, G) = 0$. To determine the asymptotic distribution of $\hat{\Delta}$ under the null decompose it as

$$\begin{aligned}
\Delta(\hat{\beta}, \hat{\gamma}, \hat{F}, \hat{G}) &= \Delta(\hat{\beta}, \hat{\gamma}, \hat{F}, \hat{G}) - \Delta(\beta^*, \gamma^*, F, G) \\
&= \Delta(\hat{\beta}, \hat{\gamma}, \hat{F}, \hat{G}) - \Delta(\hat{\beta}, \hat{\gamma}, F, G) \qquad (4) \\
&+\Delta(\hat{\beta}, \hat{\gamma}, F, G) - \Delta(\beta^*, \gamma^*, F, G) \qquad (5)
\end{aligned}$$

We refer to (4) and (5) as the Hoeffding term and the Taylor term, for reasons that will become apparent. The Hoeffding term reflects the estimation error due to the use of the estimated rather than true AUC whereas the Taylor reflects the error due to the use of estimated coefficients $\hat{\beta}, \hat{\gamma}$ rather than their probability limits β^*, γ^*. Several possibilities affecting the convergence rate arise as first-order components of the Hoeffding and Taylor terms may or may not vanish. We consider them in turn.

The Taylor term (5). Assuming $\theta(\cdot, F, G)$ is smooth enough, expand (5) in a Taylor series:

$$\begin{aligned}
\Delta(\hat{\beta}, \hat{\gamma}, F, G) &= \theta(\hat{\beta}, F, G) - \theta(\hat{\gamma}, F, G) \\
= (\hat{\beta} - \beta^*)^T \theta'(\beta^*, F, G) &+ (\hat{\beta} - \beta^*)^T \theta''(\beta^*, F, G)(\hat{\beta} - \beta^*)/2 + o_P(|\hat{\beta} - \beta^*|^2) \\
+(\hat{\gamma} - \gamma^*)^T \theta'(\gamma^*, F, G) &+ (\hat{\gamma} - \gamma^*)^T \theta''(\gamma^*, F, G)(\hat{\gamma} - \gamma^*)/2 + o_P(|\hat{\gamma} - \gamma^*|^2)
\end{aligned}$$

The primes indicate differentiation with respect to the first argument. Under the alternative $\beta \neq \gamma$, considered in Doyle-Connolly and Michael (2023), the vanishing of $\theta'(\beta^*, F, G)$ and $\theta'(\gamma^*, F, G)$ determines whether the standard estimator (DeLong et al., 1988) of the asymptotic distribution of $\hat{\Delta}$ is valid. If $\theta'(\beta^*, F, G) = \theta'(\gamma^*, F, G) = 0$, the use of estimated coefficients $\hat{\beta}, \hat{\gamma}$ does not affect the asymptotic distribution of θ or Δ, and no adjustment is necessary. When either is nonzero, an adjustment term is required, though the asymptotic distribution remains normal.

Here $\beta^* = \gamma^*$, and so $\theta'(\beta^*, F, G) = \theta'(\gamma^*, F, G)$. The first possibility is that the common value of the gradient is nonzero, in which case the normal analysis in Doyle-Connolly and Michael (2023) again applies. An example is when the data follow the conditionally normal model (15) and the coefficient estimates $\hat{\beta}, \hat{\gamma}$, have a limit β^* that is not proportional to $\Sigma^{*-1}(\mu_1 - \mu_0)$. The limit will in fact be proportional when the estimates are obtained using LDA, logistic regression, MRC, or the other examples given in Doyle-Connolly and Michael (2023), but it may not be for some other coefficient estimator or a misspecified estimator. On the other hand, if the common value is zero, (5) is of order $o(1/\sqrt{M+N})$,

$$\begin{aligned}
&\Delta(\hat{\beta}, \hat{\gamma}, F, G) \\
&= (\hat{\beta} - \beta^*)^T \theta''(F, G, \beta^*)(\hat{\beta} - \beta^*)/2 + (\hat{\gamma} - \gamma^*)^T \theta''(F, G, \gamma^*)(\hat{\gamma} - \gamma^*)/2 + o_P(|\hat{\beta} - \beta^*|^2) + o_P(|\hat{\gamma} - \gamma^*|^2). \qquad (6)
\end{aligned}$$

Complementing Doyle-Connolly and Michael (2023) this paper focuses on the second case. Doyle-Connolly and Michael (2023), Section 5.1.1, gives numerous common examples of coefficient estimation procedures, such as a well-specified logistic regression, where $\theta'(\beta^*, F, G)$ vanishes. Under the null $\beta^* = \gamma^*$, these situations imply (6).

The Hoeffding term (4). Further decompose (4) as

$$\begin{aligned} &\Delta(\hat\beta,\hat\gamma,\hat F,\hat G)-\Delta(\hat\beta,\hat\gamma,F,G)\\ =\;&\Delta(\hat\beta,\hat\gamma,\delta F,G)+\Delta(\hat\beta,\hat\gamma,F,\delta G) &(7)\\ &+\Delta(\hat\beta,\hat\gamma,\delta F,\delta G) &(8)\end{aligned}$$

where $\delta F=\hat F-F$ and $\delta G=\hat G-G$. Aside from the randomness in the coefficient estimates $\hat\beta$ and $\hat\gamma$, (7) constitutes the Hoeffding decomposition of the U-statistic $\hat\Delta$ (Hoeffiding, 1948) . Assuming $\Delta(\cdot,\cdot,\delta F,G)$ and $\Delta(\cdot,\cdot,F,\delta G)$ are differentiable ,

$$\begin{aligned} &\Delta(\hat\beta,\hat\gamma,\delta F,G)+\Delta(\hat\beta,\hat\gamma,F,\delta G)\\ =\;&(\hat\beta-\beta^*,\hat\gamma-\gamma^*)^T(\Delta'(\beta^*,\gamma^*,\delta F,G)+\Delta'(\beta^*,\gamma^*,F,\delta G))+\text{ lower order terms.} \quad (9)\end{aligned}$$

The second factor is an IID sum

$$\begin{aligned}\Delta'(\beta^*,\gamma^*,\delta F,G)+\Delta'(\beta^*,\gamma^*,F,\delta G)=&\sum_i\left(\frac{d}{d\beta}P(\beta^{*T}X_i<\beta^{*T}Y\mid X_i)-\frac{d}{d\gamma}P(\gamma^{*T}X_i<\gamma^{*T}Y\mid X_i)\right)\\ &+\sum_j\left(\frac{d}{d\beta}P(\beta^{*T}X<\beta^{*T}Y_j\mid Y_j)-\frac{d}{d\gamma}P(\gamma^{*T}X<\gamma^{*T}Y_i\mid Y_j)\right)\end{aligned} \quad (10)$$

so that, $\hat\beta$ and $\hat\gamma$ being $\sqrt{M+N}$-consistent, (7) is $O_P(1/(M+N))$.

The expression (8) is usually $o_P(1/(M+N))$ by the uniform convergence result given in Lemma 1 and asymptotically negligible, the other terms under consideration being $O_P(1/(M+N))$. The Lemma invokes concepts from empirical process theory; See Nolan and Pollard (1987) and the references there for further elaboration.

Lemma 1. *With X,Y defined as in* (2)*, suppose $h(\beta)=h(\beta,\cdot,\cdot):\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ belongs to a family of functions indexed by $\beta\in B$.*

Assuming: 1. $\{h(\beta,\cdot,\cdot):\beta\in B\}$ has an integrable envelope, i.e., $Esup_\beta h(\beta,\cdot,\cdot)<\infty$,

2. $\{h(\beta,\cdot,\cdot):\beta\in B\}$ is a VC class of functions ,

3. $\{h(\beta,X,Y):\beta\in B\}$ is a class of degenerate U-statistics, i.e., $E(h(\beta,X,Y)\mid X)=E(h(\beta,X,Y)\mid Y)=0$ for all β,

4. $M/N\to_p r\in(0,\infty)$.

Then the process

$$\beta\mapsto\frac{1}{\sqrt{MN}}\sum_{i=1}^m\sum_{j=1}^n h(\beta,X_i,Y_j) \quad (11)$$

is stochatically equicontinuous, i.e.,

$$\lim_{\epsilon\to 0}\sup E\sup_{|h(\beta)-h(\gamma)|_{L^2}<\epsilon}\left|\frac{1}{\sqrt{MN}}\sum_{i=1}^M\sum_{j=1}^N(h(\beta,X_i,Y_j)-h(\gamma,X_i,Y_j))\right|=0.$$

Proof. Rewrite h as a one-sample U-statistic on pairs of observations rather than a two-sample U-statistic on individual observations:

$$\frac{1}{\sqrt{MN}}\sum_{i=1}^{m}\sum_{j=1}^{n}h(\beta,X_i,Y_j)=\frac{1}{\sqrt{MN}}\sum_{1\le i,j\le M+N,i\neq j}h(\beta,W_i,W_j)\{D_i<D_j\}$$
$$=2\sqrt{MN}\sum_{1\le i<j\le M+N}\frac{1}{2}(h(\beta,W_i,W_j)\{D_i<D_j\}+h(\beta,W_j,W_i)\{D_j<D_i\}).$$

The bivariate function

$$(W,D),(W',D')\mapsto\frac{1}{2}(h(\beta,W,W')\{D<D'\}+h(\beta,W',W)\{D'<D\}) \tag{12}$$

is symmetric and inherits the assumed degeneracy condition from h,

$$E(h(\beta,W,W')\{D<D'\}\mid W,D)=E(h(\beta,W,W')\mid W,D,D')\{D<D'\}$$
$$=E(h(\beta,X,Y')\mid X,D=0,D'=1)\{D<D'\}=0.$$

Since VC classes are closed under pairwise sums , the function (12) also inherits the VC property from h. Nolan and Pollard (1987), Theorem 7, then gives stochastic equicontinuity of the process

$$\beta\mapsto\frac{1}{M+N}\sum_{1\le i<j\le M+N}\frac{1}{2}(h(\beta,W_i,W_j)\{D_i<D_j\}+h(\beta,W_j,W_i)\{D_j<D_i\}).$$

Since $(M+N)/\sqrt{MN}\to_p\sqrt{r}+1/\sqrt{r}\in(0,1)$, stochastic equicontinuity follows for the process (11). □

Theorem 2. *Given a sample* $(W_1,D_1),\ldots,(W_{M+N},D_{M+N})$, *from* (1) *and coefficient estimates* $\hat\beta,\hat\gamma$, *assume*

1. $\beta^*=\gamma^*$,
2. *Influence functions are available for* $\hat\beta$ *and* $\hat\gamma$, *i.e., square-integrable functions* $\psi_{\hat\beta},\psi_{\hat\gamma}$ *such that* $\hat\beta-\beta^*=\sum_{i=1}^{M}\psi_{\hat\beta}(W_i,D_i)+o_P(N^{-1/2})$ *and* $\hat\gamma-\gamma^*=\sum_{i=1}^{M}\psi_{\hat\gamma}(W_i,D_i)+o_P(N^{-1/2})$,
3. $\theta(\cdot,F,G)$ *is twice differentiable at* β^*, *and the first derivative vanishes there,*
4. *The term* $\Delta(\cdot,\cdot,\delta F,G)+\Delta(\cdot,\cdot,F,\delta G)$ *is differentiable at* (β^*,γ^*).

Then, the asymptotic distribution of $\Delta(\hat\beta,\hat\gamma,\hat F,\hat G)$ *is the distribution of*

$$a^Tb+a^T\begin{pmatrix}\theta''(F,G,\beta^*) & 0\\ 0 & -\theta''(F,G,\gamma^*)\end{pmatrix}a/2, \tag{13}$$

where (a,b) *is mean-zero multivariate normal with*

$$a\sim\lim\sqrt{M+N}(\hat\beta-\beta^*,\hat\gamma-\gamma^*),b\sim\lim\sqrt{M+N}\left(\Delta'(\beta^*,\gamma^*,F,\delta G)+\Delta'(\beta^*,\gamma^*,\delta F,G)\right).$$

Proof. We first show that $\Delta(\hat\beta,\hat\gamma,\delta F,\delta G)=o(1/N)$. Define $\theta_0(\beta,x,y):(\beta,x,y)\mapsto\psi(\beta^Tx,\beta^Ty)-$

$E(\psi(\beta^TX,\beta^TY)\mid X=x)-E(\psi(\beta^TX,\beta^TY)\mid Y=y)+E\psi(\beta^TX,\beta^TY)$. The class of functions $\{\theta_0(\beta,\cdot,\cdot):\beta\}$ is degenerate in the sense of Lemma 1 and a VC class of functions (see (Sherman, 1993), Corollary to Theorem 4). By Lemma 1, the process mapping β to

$$\sqrt{MN}\int\psi(\beta^Tx,\beta^Ty)d(\hat F-F)(x)d(\hat G-G)(y)=\frac{1}{\sqrt{MN}}\sum_{i=1}^{m}\sum_{j=1}^{n}\theta_0(\beta,X_i,Y_j)$$

is stochastically equicontinuous.

Given $\epsilon'>0$, there is $\epsilon(\epsilon')>0$ given by bounded convergence such that $|\theta_0(\beta,\cdot,\cdot)-\theta_0(\gamma,\cdot,\cdot)|_2<\epsilon(\epsilon')$ whenever $|\beta-\gamma|<\epsilon'$. Then,

$$\begin{aligned}&P\Big((M+N)\Delta(\hat\beta,\hat\gamma,\delta F,\delta G)>\epsilon(\epsilon')\Big)\\&\le P(|\hat\beta-\hat\gamma|>\epsilon')+P\left(\sup_{|\theta_0(\beta,\cdot,\cdot)-\theta_0(\gamma,\cdot,\cdot)|_2<\epsilon(\epsilon')}\left|\frac{1}{\sqrt{MN}}\sum_{i=1}^{M}\sum_{j=1}^{N}(\theta_0(\beta,X_i,Y_j)-\theta_0(\gamma,X_i,Y_j))\right|>\epsilon(\epsilon')\right)+o(1)\\&=o(1).\end{aligned}$$

Since $\hat\beta,\hat\gamma$ are $\sqrt{N}$-consistent, the Hoeffding term $\Delta(\cdot,\cdot,\delta F,G)+\Delta(\cdot,\cdot,F,\delta G)$ is differentiable at (β^*,γ^*), and the AUC gradient $\theta'(\cdot,F,G)$ vanishes there,

$$\begin{aligned}\Delta(\hat\beta,\hat\gamma,\hat F,\hat G)&=\Delta(\hat\beta,\hat\gamma,F,\delta G)+\Delta(\hat\beta,\hat\gamma,\delta F,G)+\Delta(\hat\beta,\hat\gamma,\delta F,\delta G)+\Delta(\hat\beta,\hat\gamma,F,G)\\&=(\hat\beta-\beta^*,\hat\gamma-\gamma^*)^T(\Delta'(\beta^*,\gamma^*,F,\delta G)+\Delta'(\beta^*,\gamma^*,\delta F,G))\\&+\begin{pmatrix}\hat\beta-\beta^*\\\hat\gamma-\gamma^*\end{pmatrix}^T\begin{pmatrix}\theta''(F,G,\beta^*)&0\\0&-\theta''(F,G,\gamma^*)\end{pmatrix}\begin{pmatrix}\hat\beta-\beta^*\\\hat\gamma-\gamma^*\end{pmatrix}+o_P(1/N).\end{aligned}\tag{14}$$

The non-negligible part of $\Delta(\hat\beta,\hat\gamma,\hat F,\hat G)$ is a continuous function of

$$(\psi_{\hat\beta},\psi_{\hat\gamma},\Delta'(\beta^*,\gamma^*,F,\delta G),\Delta'(\beta^*,\gamma^*,\delta F,G))$$

the components of which are IID sums of terms with finite variances and jointly asymptotically normal by the CLT.

□

We briefly comment on the assumptions of Theorem 2. The assumption that $\beta^*=\gamma^*$ is how we have interpreted the null hypothesis $\Delta=0$, as have others, e.g., Heller et al. (2017). Influence functions are available for many common coefficient estimation procedures, though assumption 2 does rule out coefficient estimation procedures that converge more slowly than the parametric rate. Assumption 3 is more consequential. As noted earlier, the vanishing or not of the gradient of the AUC determines whether a $\sqrt{M+N}$ or $M+N$ normalization is appropriate. The final assumption is the differentiability of $\Delta(\cdot,\cdot,\delta F,G)+\Delta(\cdot,\cdot,F,\delta G)$, which shows up as Assumption A4(i) in Sherman (1993) and is implicitly assumed in Theorem

1 of Heller et al. (2017). A simple sufficient condition is the existence of densities for X and Y. Further conditions are discussed in Section 8 of Sherman (1993).

Finally, we compare the result presented here for the null hypothesis $\beta^* = \gamma^*$ to the result presented in Doyle-Connolly and Michael (2023) for the alternative hypothesis $\beta^* \neq \gamma^*$.

1. *The method.* In the case of the alternative $\beta^* \neq \gamma^*$ the approach taken in Doyle-Connolly and Michael (2023) was to form first-order expansions of the AUC, and an expansion for the difference of AUCs followed as a corollary. The analogous approach here, where the first-order terms vanish, would be to form second-order expansions of the AUC and then take the difference. However, the behavior of this expansion is difficult to ascertain because of the interaction between the "mixed partials" $(\hat{\beta} - \beta^*)^T \theta'(\beta^*, \delta F, G)$, etc., in (9), and the quadratic Hoeffding term $\theta(\hat{\beta}, \delta F, \delta G)$ in (8). The approach taken directly targets the difference of the AUCs. It is less general but simpler since the difference of the quadratic Hoeffding terms vanishes by Lemma 1.

2. *Estimation.* Estimation of the asymptotic distribution (13) is substantially more complicated than the distribution under the alternative. First, the Hessian rather than the gradient of the AUC must be estimated. Numerical methods were used in Doyle-Connolly and Michael (2023). Second, under the null the asymptotic variance of the Hoeffding term, denoted b in Theorem 2, and its covariance with the influence functions, a in Theorem 2, must be estimated.

4 Estimation

Specification of the asymptotic null distribution (13) in general requires

1. the Hessian of the AUC at β^*,

2. an influence function for the coefficient estimation procedure, and

3. the variance matrix for the combined influence functions for $\hat{\beta}$ and $\hat{\gamma}$ and the Hoeffding gradient (4)

   (a) the variance of the Hoeffding gradient, and

   (b) the variance matrix of the influence functions

   (c) the matrix of covariances between the Hoeffding gradient (4) and the coefficient estimates.

The Hessian of the AUC and the variance matrix of the Hoeffding gradient are common to any model with the same covariate distribution. The influence function and moments involving the coefficient estimates must then be obtained based on the chosen coefficient estimation procedure. Below we describe parametric and non-parametric estimation of these quantities.

4.1 Parametric estimation

We compute the asymptotic distribution for $\hat{\Delta}$ given by Theorem 2 assuming that the covariates (1) belong to a parametric family. Although parametric assumptions are too strong for many applications, this analysis will show the types of quantities that need to be estimated in the non-parametric setting, as well as provide a benchmark for non-parametric estimation in the simulations below.

We model the covariates in each class as normally distributed,

$$\begin{aligned} W|D = d \sim F_d = N_p(\mu_d, \Sigma_d), \Sigma_d > 0, d \in \{0,1\}, \mu_1 \neq \mu_0 \\ P(D=1) = 1 - P(D=0) = \pi_1. \end{aligned} \tag{15}$$

We use the normal model for two principal reasons. The first is that it leads to closed form expressions for many of the quantities involved in the asymptotic distribution. The second is that it contains a sub-model that may be viewed as a well-specified LDA model or well-specified logistic regression model, the latter of which belongs to the class of generalized regression models proposed by Han (1987). The last point makes the normal model particularly suitable for this study since the two models mentioned are the two cases of parametric approaches to testing $H_0 : \Delta = 0$ known to us from the literature: 1) normal covariates and LDA coefficients (Demler et al., 2011) and 2) unspecified covariates with the MRC estimator for the coefficients (Heller et al., 2017). These are re-analzyed below in Section 4.1.2 for MRC and 5 for LDA. In both, it is supposed that the coefficients are nested. That is, $\hat{\gamma}$ is formed based on the full vector of covariates W whereas $\hat{\beta}$ is based on a subset, while the null hypothesis $\beta^* = \gamma^*$ holds for the probability limits $\beta^* = \lim \hat{\beta}, \gamma^* = \lim \hat{\gamma}$. This setup is commonly used for testing if there is any improvement in discrimination provided by the additional covariates.

Formulas for the conditionally normal model (15) are given in Proposition 3, corresponding to items 1 and 3a in the list of parameters to be estimated. The remaining parameters involve the coefficient estimation procedure and are given in Propositions 4 and 6 below for LDA and MRC, respectively.

Proposition 3. *Under the conditionally gaussian model* (15),

1. *The Hessian of the AUC is* $\frac{d^2}{d\beta^2}\theta(\beta, F, G) = u''(\beta, \Sigma_0 + \Sigma_1, \mu)$, *where, for* $\beta \in R^p, w \in R^p, \Sigma \in R^{p \times p}$ $u : (\beta, \Sigma, w) \mapsto \Phi(\beta^T w / \sqrt{\beta^T \Sigma \beta})$. *At* $\beta = c\Sigma^{-1}\mu, c \in \mathbb{R}, c \neq 0$, *the Hessian takes the form*

$$\frac{d^2}{d\beta^2} P(\beta^T X < \beta^T Y)\bigg|_{\beta = c\Sigma^{-1}\mu} = \phi(c^{-1}\sqrt{q})c^{-1}q^{-1/2}(q^{-1}\Sigma\beta\beta^T\Sigma - \Sigma).$$

2. *The Hoeffding gradient* $\frac{d}{d\beta}(\theta(\beta, F, \delta G) - \theta(\beta, \delta F, G))$ *is*

$$N^{-1}\sum_{i=1}^{N} u'(\beta, \Sigma_0, Y_i - \mu_0) + M^{-1}\sum_{i=1}^{M} u'(\beta, \Sigma_1, \mu_1 - X_i) - 2u'(\beta, \mu, \Sigma)$$

with u as above, and its variance is given by

$$N^{-1}J_3(\beta,\mu,\Sigma_1,\Sigma_0)+M^{-1}J_3(\beta,\mu,\Sigma_0,\Sigma_1)-(M^{-1}+N^{-1})u'(\beta,\mu,\Sigma)^{\otimes 2},$$

where for $d\in\{0,1\}$,

$$J_3:(\beta,\mu,\Sigma_{1-d},\Sigma_d)\mapsto\left(J_1/2+(\beta^TJ_1\beta)/(2q_d^2)\Sigma_d\beta\beta^T\Sigma_d-(\Sigma_d\beta\beta^TJ_1+J_1\beta\beta^T\Sigma_d)/(2q_d)\right)/\sqrt{2\pi},$$

and $J_1=J_1(\beta,\sqrt{2q_d^{-1}}\mu,2q_d^{-1}\Sigma_{1-d}),q_d=\beta^T\Sigma_d\beta,q=\beta^T(\Sigma_0+\Sigma_1)\beta,$ where

$$J_1:(\beta,\mu,\Sigma)\mapsto\phi(\frac{\beta^T\mu}{\sqrt{1+q}})\frac{1}{\sqrt{1+q}}((\mu-\frac{\beta^T\mu}{1+q}\Sigma\beta)(\mu^T-\frac{\beta^T\mu}{1+q}\beta^T\Sigma)+\Sigma-\frac{(\Sigma\beta)^{\otimes 2}}{1+q}).$$

Proof. These formulas can be demonstrated using standard manipulations with the gaussian PDF and CDF. The quantities J_1 and J_3 defined above carry the interpretations

$$J_1(\beta,\mu,\Sigma)=E(\phi(\beta^TW)WW^T)\text{ for }W\sim N(\mu,\Sigma)$$
$$J_3(\beta,\mu,\Sigma_0,\Sigma_1)=E(u'(\beta,\Sigma_1,W)^{\otimes 2})\text{ for }W\sim N(\mu,\Sigma_0).$$

□

4.1.1 LDA for coefficient estimation

Linear discriminant analysis builds a rule that classifies a new sample w as control or case based on the sign of $\hat\beta_{LDA}^Tw$. The coefficients are computed as

$$\hat\beta_{LDA}=\hat\Sigma^{-1}(\hat\mu_1-\hat\mu_0),\text{ where}$$
$$\hat\mu_1-\hat\mu_0=N^{-1}\sum_i Y_i-M^{-1}\sum_i X_i$$
$$\hat\Sigma=(M+N)^{-1}\left(\sum_i(X_i-\hat\mu_0)(X_i-\hat\mu_0)^T+\sum_i(Y_i-\hat\mu_1)(Y_i-\hat\mu_1)^T\right).$$

An intercept is usually computed when carrying out LDA but may be ignored here since the AUC does not change when both classes undergo a common shift. The LDA parameter estimates tend in probability to

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

These probability limits do not lead to the optimal classification rule for normal data unless the class covariances Σ_0 and Σ_1 are equal, which is a usual assumption of LDA. However, the conclusion of Theorem 2 does not depend on the correct specification of the coefficient model. The conclusion only depends on the asymptotic variance at whatever the probability limit of the estimates may be.

Proposition 4 gives formulas for the terms in (13) that are specific to the LDA procedure. In comparing nested that are nested under the null, β^* is a proper subset of the components of γ^*. Subscripts "β" and "γ" on vectors indicate the vector formed from the restricted or full set of components. Likewise, "$\beta\beta$", "$\beta\gamma$", etc. subscripts on matrices subset the corresponding rows and columns.

Proposition 4. *1. Influence functions for the LDA coefficient estimates based on full and restricted samples are given by*

$$\psi_{\hat\gamma} : (w,d) \mapsto \sum_{i=0,1} \pi_i^{-1}\{d=i\} \left(\Sigma^{*-1}w + \pi_i \Sigma^{*-1}(w-\mu_i)^{\otimes 2}\Sigma^{*-1}\mu)\right)$$
$$\psi_{\hat\beta} : (w,d) \mapsto (\Sigma_{\beta\beta})^{-1} I_{\beta\gamma}\psi_{\hat\gamma}(w,d),$$

with variances under model (15) *given by*

$$\begin{aligned}\mathrm{Var}(\psi_{\hat\gamma}(W,D)) &= \Sigma^{*-1}(\Sigma_0/\pi_0 + \Sigma_1/\pi_1)\Sigma^{*-1}\\ &\quad + \Sigma^{*-1}\left(\sum_{d=0,1}(\pi_d(\Sigma_d\beta^*\beta^{*T}\Sigma_d + (\beta^{*T}\Sigma_d\beta^*)\Sigma_d))\right)\Sigma^{*-1}\\ \mathrm{Var}(\psi_{\hat\beta}(W,D)) &= (\Sigma_{\beta\beta})^{-1}I_\beta \,\mathrm{Var}(\psi_{\hat\gamma}(W,D)) I_\beta^T(\Sigma_{\beta\beta})^{-1}\\ \mathrm{Cov}(\psi_{\hat\beta}(W,D),\psi_{\hat\gamma}(W,D)) &= (\Sigma_{\beta\beta})^{-1}I_\beta\,\mathrm{Var}(\psi_{\hat\gamma}(W,D)).\end{aligned}$$

2. The covariance between the influence function for the coefficient estimate under model (15) *and the Hoeffding gradient at β^* is*

$$\begin{aligned}&\mathrm{Cov}(\psi_{\hat\gamma}(W,D), \frac{d}{d\beta}(\theta(\beta^*,F,\delta G) - \theta(\beta^*,\delta F, G)))\\ &= \sum_{d=0,1}\sqrt{q_d}\Sigma^{*-1}(I+\pi_d(\mu\beta^{*T}+\mu^T\beta^* I))\left(J_1(\beta^*,\frac{\mu}{\sqrt{q_{1-d}}},\frac{\Sigma_d}{q_{1-d}}) - \frac{\mu}{\sqrt{q_{1-d}}}J_4'(d)\right)(\beta^*\beta^{*T}\frac{\Sigma_{1-d}}{q_{1-d}} - I)\\ &- \sum_{d=0,1}\pi_d\Sigma^{*-1}q_{1-d}\left(J_4''(d) - (\frac{\Sigma_d}{q_{1-d}} + \mu^{\otimes 2}q_{1-d})\beta J_4'(d)\right)(\beta^*\beta^{*T}\frac{\Sigma_{1-d}}{q_d} - I),\\ &\mathrm{Cov}(\psi_{\hat\beta}(W,D)), \frac{d}{d\beta}(\theta(\beta^*,F,\delta G) - \theta(\beta^*,\delta F,G)))\\ &= (\Sigma_{\beta\beta})^{-1}I_\beta\,\mathrm{Cov}(\psi_{\hat\gamma}(W,D), \frac{d}{d\beta}(\theta(\beta^*,F,\delta G)-\theta(\beta^*,\delta F,G))),\end{aligned}$$

where J_1 is as in Proposition 3, $J_4 : (\beta,\mu,\Sigma) \mapsto \Phi(\beta^T\mu/\sqrt{1+\beta^T\Sigma\beta})$, primes indicate differentiation with respect to β, and for $d \in \{0,1\}$, $J_4(d), J_4'(d), J_4''(d)$ above are evaluated at $(\beta^*, \frac{\mu}{\sqrt{q_{1-d}}}, \frac{\Sigma_d}{q_{1-d}})$.

4.1.2 MRC for coefficient estimation

When the class covariance matrices in (15) are equal, $\Sigma_0 = \Sigma_1$, the conditionally normal model is also a logistic model,

$$\begin{aligned}P(D=1 \mid W) = &\frac{\pi_1 G'(w)}{\pi_0 F'(w) + \pi_1 G'(w)}\\ &= \mathrm{expit}(\log(\pi_0/\pi_1) + (\mu_0\Sigma\mu_0 - \mu_1\Sigma\mu_1)/2 - \beta_{LDA}^T w)),\end{aligned}$$

with β_{LDA} defined as in Section 4.1.1. The logistic model belongs to a larger family of parametric regression models proposed by Han (1987). These models relate a vector of covariates W to a response D, which need not be binary. As an estimator for the parameter Han proposed the maximum rank correlation. When D is in fact binary, as here, the MRC takes the form

$$\hat{\beta}_{MRC} = \underset{\beta:\beta_1=1}{\arg\max}\, \theta(\beta, \hat{F}, \hat{G}) \tag{16}$$

That is, the estimator maximizes the empirical AUC subject to its first component being 1. Without some type of normalization the maximization isn't well-posed as $P(\beta^T X < \beta^T Y) = P(c\beta^T X < c\beta^T Y), c > 0$, and we follow Heller et al. (2017) and others in setting the first component to be 1. Han (1987); Sherman (1993) gives conditions under which the MRC is consistent, i.e., $\beta^* = \beta_{LDA}$, and aymptotically normal. For the conditionally normal model (15) consistency follows from Lemma 5.

Lemma 5. *In the normal model* (15) *with* $\Sigma = \Sigma_0 = \Sigma_1$

1. *The only stationary points of the AUC* $\theta(\beta, F, G)$ *are given by* $\beta \propto \Sigma^{-1}(\mu_1 - \mu_0)$, *and there is a unique stationary point* β^* *such that the first component is equal to* 1, *permuting the components of* β^* *if necessary.*

2. *The Hessian of the AUC at a stationary point* β^* *with first component equal to* 1, $\theta''(\beta^*, F, G)$, *after removing the first row and column, is negative definite, switching the class labels if necessary.*

Proof. 1. The gradient of the AUC

$$\frac{d}{d\beta} u(\beta, \Sigma, \mu) = \phi\left(\frac{\beta^T \mu}{\sqrt{\beta^T \Sigma \beta}}\right) \frac{1}{\sqrt{\beta^T \Sigma \beta}} \left(I - \frac{\Sigma \beta \beta^T}{\beta^T \Sigma \beta}\right) \mu$$

is equal to 0 iff μ lies in the nullspace of $I - \Sigma\beta\beta^T / \beta^T \Sigma \beta$. An example of such is $\mu \propto \Sigma\beta$ and since $I - \Sigma\beta\beta^T/\beta^T\Sigma\beta$ has rank $\geq p - 1$ this is the only example. Let $\beta^* = c\Sigma^{-1}\mu$ with $c \in \mathbb{R}, c > 0$, chosen so that the first component of β^* is 1. Such a c can be found by switching the class labels if all components of β^* are ≤ 0, and then by permuting the indices of β^* so that the first component is > 0, if it isn't already.

At β^*, $\theta''(\beta^*, F, G)$ takes the form given in Prop. 3. Given $x \in \mathbb{R}^p$, by the Cauchy-Schwarz inequality,

$$x^t \theta''(\beta^*, F, G) x = \phi(\sqrt{\beta^{*T} \Sigma \beta^*}/c)/c/(\beta^{*T}\Sigma\beta^*)^{3/2}((\beta^{*T}\Sigma x)^2 - (\beta^{*T}\Sigma\beta^*)(x^t \Sigma x)) \leq 0$$

and $= 0$ iff $x \propto \beta^*$. Then $\theta''(\beta^*, F, G)$ with the first row and column removed must be negative definite as otherwise the nullspace contains a vector with first component equal to 0, which cannot be proportional to β^* given the requirement that the latter has first component equal to 1. □

Corresponding to Proposition 4, Proposition 6 gives formulas for the terms in the asymptotic distribution (13) for $\hat{\Delta}$ that involve the MRC coefficient estimates.

Proposition 6. *1. An influence functions for the MRC estimate $\hat{\beta}$ is given by*

$$\psi_{\hat{\beta}} : (w,d)$$
$$\mapsto -\theta''(\beta^*,F,G)^{-1}(\pi_0^{-1}\frac{d}{d\beta}E(\psi(\beta^T w,\beta^T Y)\mid w)\{d=0\}+\pi_1^{-1}\frac{d}{d\beta}E(\psi(\beta^T X,\beta^T w)\mid w)\{d=1\}),$$

with variance matrix under the normal model (15)

$$\mathrm{Var}(\psi_{\hat{\beta}}(W,D))=\theta''(\beta^*,F,G)^{-1}\,\mathrm{Var}(\theta'(\beta^*,F,\delta G)+\theta'(\beta^*,\delta F,G))\theta''(\beta^*,F,G)^{-1}.$$

Formulas for the above quantities are given in Proposition 3. Analogous expression for the influence function of $\hat{\gamma}$ and its covariance matrix by substituting above $\hat{\gamma}$ for $\hat{\beta}$, γ^* for β^*. The asymptotic covariance between the influence functions for $\hat{\beta}$ and $\hat{\gamma}$ is given by

$$\mathrm{Cov}(\psi_{\hat{\beta}}(W,D),\psi_{\hat{\gamma}}(W,D))=\theta''(\beta^*,F_\beta,G_\beta)^{-1}\,\mathrm{Var}(\theta'(\beta^*,F,\delta G)+\theta'(\beta^*,\delta F,G))_{\beta\gamma}\theta''(\gamma^*,F_\gamma,G_\gamma)^{-1}$$
$$-\theta''(\beta^*,F_\beta,G_\beta)^{-1}u'(\beta^*,\Sigma_\beta,\mu_\beta/\sqrt{2})u'(\gamma^*,\Sigma_\gamma,\mu_\gamma/\sqrt{2})^T\theta''(\gamma^*,F_\gamma,G_\gamma)^{-1}.$$

2. The covariances between the influence functions and the Hoeffding gradient at β^* are given by

$$\mathrm{Cov}(\psi_{\hat{\beta}}(W,D),\frac{d}{d\beta}(\theta(\beta^*,F,\delta G)-\theta(\beta^*,\delta F,G)))=\mathrm{Cov}(\psi_{\hat{\beta}}(W,D),\psi_{\hat{\gamma}}(W,D))\theta''(\gamma^*,F_\gamma,G_\gamma)$$

$$\mathrm{Cov}(\psi_{\hat{\gamma}}(W,D),\frac{d}{d\beta}(\theta(\beta^*,F,\delta G)-\theta(\beta^*,\delta F,G)))=\mathrm{Var}(\psi_{\hat{\gamma}}(W,D))\theta''(\gamma^*,F_\gamma,G_\gamma).$$

Proof. 1. From (14),

$$\theta(\beta,\hat{F},\hat{G})-\theta(\beta^*,\hat{F},\hat{G})$$
$$=\frac{1}{2}(\beta-\beta^*)^T\theta''(\beta^*,F,G)(\beta-\beta^*)+(\beta-\beta^*)^t(\theta'(\beta^*,F,\delta G)+\theta'(\beta^*,G,\delta F))+o_P(1/N).$$

Viewing the above expressions as functions of β, Sherman (1993), Theorem 2, asserts that the maximizer of the left-hand side, i.e, $\hat{\beta}$, is within $o_P(1/\sqrt{N})$ of the maximizer of the approximating quadratic on the right-hand side, i.e., $\beta^*-\theta''(\beta^*,F,G)^{-1}(\theta'(\beta^*,F,\delta G)+\theta'(\beta^*,G,\delta F))$, which is to say

$$\hat{\beta}-\beta^*+\theta''(\beta^*,F,G)^{-1}(\theta'(\beta^*,F,\delta G)+\theta'(\beta^*,G,\delta F))=o_P(1/\sqrt{N}).$$

The assumptions of the cited theorem are verified by Lemma 5. The other sub-parts are straightforward, e.g., the formula for the covariance follows as $u'(\beta^*_\beta,\Sigma_\beta,w_\beta)$ is just the first p_β rows of $u'(\beta^*_\gamma,\Sigma_\gamma,w_\gamma)$.

□

As the maximizer of the index AUC, there is a sense in which the MRC (16) is more closely related to the index AUC than other coefficient estimators such as LDA and logistic regression. It turns out that the error in $\hat{\Delta}$ due using to the empirical rather than true AUC (4) is asymptotically the same error due to the use of MRC estimates (5) rather than their probability limit. This fact can be derived from the form of the influence function given by Prop. 6 and does not rely on the normality assumption (15). Thus we recover a simplified expression for the asymptotic distribution of Δ whenever the coefficient estimation procedure is MRC, previously given by Heller et al. (2017). The calculation given here avoids the cited proof's appeal to an expansion justified only heuristically and then only in the case that the coefficient estimate is an MLE.

Corollary 7 (Heller et al. (2017))**.** *Suppose the assumptions of Theorem 2 hold, and further that the MRC estimator* (16) *is* $\sqrt{M+N}$*-consistent for the maximizer of the AUC. Then the asymptotic distribution of* $\hat{\Delta}(\hat{\beta}, \hat{\gamma}, \hat{F}, \hat{G})$ *is the distribution of* $\frac{1}{2}a^T b$*, with* a *and* b *defined as in Theorem 2; equivalently, the distribution of* $\sum \lambda_i \chi_i^2$ *where* χ_i *are independent chi-squared random variables and* λ_i *are the eigenvalues of* $((\theta''^{-1})_{\gamma/\beta})^{-1}(\lim \mathrm{Var}(\sqrt{N}\hat{\gamma}))_{\gamma/\beta,\gamma/\beta}$.

Proof. By Prop. 6,

$$\begin{aligned} a = \lim \sqrt{N}(\hat{\beta} - \beta^*, \hat{\gamma} - \gamma^*) &= -\lim \sqrt{N}((\theta''_{\beta\beta})^{-1}\nabla_\beta, (\theta'')^{-1}\nabla) \\ b &= \lim \sqrt{N}(\nabla_\beta, -\nabla) \\ a^T b = \lim -n({\nabla_\beta}^T(\theta''_{\beta\beta})^{-1}\nabla_\beta - \nabla(\theta'')^{-1}\nabla) &= -a^T \begin{pmatrix} \theta'' & 0 \\ 0 & -\theta' \end{pmatrix} a \end{aligned}$$

so $a^T b + \frac{1}{2}a^T \begin{pmatrix} \theta'' & 0 \\ 0 & -\theta' \end{pmatrix} a = \frac{1}{2}a^T b = -\frac{1}{2}\left({\nabla_\beta}^T(\theta''_{\beta\beta})^{-1}\nabla_\beta - \nabla^T(\theta'')^{-1}\nabla\right)$. Letting $S = D - B^T A^{-1} B$ denote the Schur complement of A in θ'', the last expression may be written as

$$-\frac{1}{2}|S^{-1/2}(B^T A^{-1}\nabla_\beta - \nabla_{\gamma/\beta})|^2. \tag{17}$$

Finally, since $\nabla = -\theta''(\hat{\gamma}-\gamma^*)+o_P(N^{-1/2})$ is asymptotically normal with variance $\theta'' \lim \mathrm{Var}(\sqrt{N}\hat{\gamma})\theta''$, $B^T A^{-1}\nabla_\beta - \nabla_{\gamma/\beta}$ is asymptotically normal with variance

$$S(\lim \mathrm{Var}(\sqrt{N}\hat{\gamma}))_{\gamma/\beta,\gamma/\beta} S = ((\theta''^{-1})_{\gamma/\beta})^{-1}(\lim \mathrm{Var}(\sqrt{N}\hat{\gamma}))_{\gamma/\beta,\gamma/\beta}((\theta''^{-1})_{\gamma/\beta})^{-1}. \tag{18}$$

It follows after diagonalizing the variance that the quadratic form (18) is the same as the stated combination of chi-squared random variables. □

By (17) of the proof, the asymptotic distribution of $\hat{\Delta}$ is non-positive. That is, with MRC the difference of AUCs can only increase or stay the same with additional covariates.

4.2 Non-parametric estimation

We next describe an approach to non-parametrically estimate the asymptotic distribution (13) of the difference in AUCs. Suppose we are given a sample of covariates and binary statuses, $(W_1, D_1), \ldots, (W_{M+N}, D_{M+N})$, and a coefficient estimation procedure, in the form of influence functions $\psi_{\hat{\beta}}$ and $\psi_{\hat{\gamma}}$. Based on this sample we estimate (13) using numerical derivatives and empirical moments.

The asymptotic variance of $\hat{\beta}$ and $\hat{\gamma}$ may be approximated from the influence function, i.e., the empirical variance of

$$\psi_{\beta}(W_1, D_1), \ldots, \psi_{\beta}(W_{M+N}, D_{M+N}). \tag{19}$$

Terms in the Hoeffding gradient (10)

$$\begin{aligned} &\frac{d}{d\beta} P(\beta^{*T} X_i < \beta^{*T} Y \mid X_i), i = 1, \ldots, M, \\ &\frac{d}{d\beta} P(\beta^{*T} X < \beta^{*T} Y_j \mid Y_j), j = 1, \ldots, N \end{aligned}$$

may be approximated by numerically differentiating

$$\begin{aligned} \beta \mapsto P_{\hat{G}}(\beta^T X_i < \beta^T \cdot) &= N^{-1} \sum_{j=1}^{N} \{\beta^T X_i < \beta^T Y_j\}, i = 1, \ldots, M, \\ \beta \mapsto P_{\hat{F}}(\beta^T \cdot < \beta^T Y_j) &= M^{-1} \sum_{i=1}^{M} \{\beta^T X_i < \beta^T Y_j\}, j = 1, \ldots, N \end{aligned} \tag{20}$$

at $\beta = \hat{\beta}$. The gradient at γ^* may be estimated likewise. The variance of the Hoeffding gradient is estimated as the empirical variance of (20). The covariance of the Hoeffding gradient and coefficient estimates is estimated as the empirical covariance between (19) and the numerical derivatives of (20). Finally, the Hessian of the AUC is estimated numerically using the empirical AUC at $\hat{\beta}$.

The simulations below use a simple finite difference approximation for the Hessian and the terms of the Hoeffding gradient. For example, the i, j component of

$$\frac{d^2}{d\beta^2} \theta(\beta, F, G)|_{\beta=\beta^*}$$

is estimated by

$$\begin{aligned} (2\epsilon)^{-2}(\theta(\hat{\beta} + \epsilon e_i + \epsilon e_j, \hat{F}, \hat{G}) + \theta(\hat{\beta} - \epsilon e_i + \epsilon e_j, \hat{F}, \hat{G}) + \\ \theta(\hat{\beta} + \epsilon e_i - \epsilon e_j, \hat{F}, \hat{G}) + \theta(\hat{\beta} - \epsilon e_i - \epsilon e_j, \hat{F}, \hat{G})), \end{aligned} \tag{21}$$

where e_i, e_j are standard basis vectors in $\mathbb{R}^p$ and ϵ is small in magnitude. Since (21) is a step function, the step size ϵ used in the finite difference approximation must be chosen with care. A plot of the empirical AUC along lines through $\hat{\beta}$ helps in determining the appropriate scale. Similar remarks apply to computation of the numerical gradients in (20). More quantitative guidance is given in Section 7 of Sherman (1993), which discusses non-parametric estimation of similar quantities, or Chay and Honore (1998), which applies these methods to a data set. The performance of the estimates is examined below using synthetic data.

5 Simulation

We examine the finite-sample performance of the parametric and non-parametric estimators, first checking empirical CDFs and then turning to error rates of tests based on the estimators. We consider two methods of coefficient estimation, LDA and logistic regression.

The normal covariate formulas given in Proposition 3 and the LDA coefficient estimation formulas given in Proposition 4 can be substituted into Theorem 2's formula for the asymptotic distribution of $\hat{\Delta}$. Assumption 4 follows from Prop. 3, Assumption 2 from Prop 6, and Assumption 3 from Lemma 5. The result is an "oracle estimator" for the asymptotic distribution, giving a benchmark for the performance of actual estimators and isolating the effect of the asymptotic distributional approximation from the estimation of nuisance parameters. A parametric estimator is obtained by substituting $\hat{\beta}$ for β^*, and empirical estimates for μ_d and Σ_d, $d = 0, 1$. Finally, a non-parametric estimator is obtained under the approach of Section 4.2.

Convenient expressions such as given in Section 4.1.1 for LDA are not available for logistic regression, so only the non-parametric estimator from among the proposed estimators is presented here. Logistic regression is included because of its popularity in applied work and because past literature has frequently noted the lack of a valid test of the null $H_0 : \Delta = 0$, e.g., Heller et al. (2017); Lee (2021).

5.1 Distributions

Figure 1a plots the empirical CDF for a synthetically generated sample of the difference in AUCs using LDA, along with the oracle, parametric, and non-parametric estimates of the CDF. The total number of controls and cases are 50 (left panel) and 150 (right panel), in a 2:1 ratio in both panels. The parametric and non-parametric estimates are formed as an average of estimated CDFs based on 500 samples. In Figure 1b the empirical CDF for a sample using logistic regression is plotted along with an average of non-parametric estimates of the CDF. A 95% pointwise confidence band based on the sampled CDFs is also given.

5.2 Error rates

Under the LDA model of Section 4.1.1, the null hypothesis of equality of the full and reduced model AUCs is, by Prop. 3,

$$H_0 : \Phi(\beta^{*T}(\mu_1 - \mu_0)/\sqrt{\beta^{*T}(\Sigma_0 + \Sigma_1)\beta^*}) = \Phi(\gamma^{*T}(\mu_1 - \mu_0)/\sqrt{\gamma^{*T}(\Sigma_0 + \Sigma_1)\gamma^*})$$

which is the same as equality of the Mahalanobis distances between the indexes for the control and case classes in the full and reduced models. As Demler et al. (2011) point out, when the class covariances are equal, $\Sigma_0 = \Sigma_1$, the scaled difference of the Mahalonobis distances follows an F-distribution under the null of no difference. In Figure 2a the rejection rate of a nominal 95% test of $H_0 : \Delta = 0$ based on the F-statistic is compared with a test based on the asymptotic CIs given by Theorem 2. Under the null, β^* has 4 non-zero components γ^* has 2 additional components set to 0. The alternatives are formed by increasing the common value of these 2 components. The F-test and the tests baesd on the oracle and non-parametric estimators perform best. Somewhat suprisingly, the test based on the parametric estimator has difficulty controlling the false positive rate. As this estimator differs from the oracle estimator only in using estimates of simple nuisance parameters, this poor performance improves with sample size. The Delong test has low power which does not improve with increased sample size. The poor power of the Delong test prompted initial research into valid tests for the difference in AUCs based on fitted values; compare Figure 3 of Demler et al. (2012).

Fig. 2b presents the error rate analysis for logistic regression. The non-parametric estimator's FPR is slightly higher than the nominal rate at $M + N = 50$ but is controlled at $M + N = 150$. As with LDA, the Delong test has poor power which cannot improve much with sample size as it is based on an incorrect asymptotic distribution and rate of convergence.

6 Discussion

We have presented the asymptotic null distribution of the difference of index AUCs under general conditions. This analysis complements the simpler analysis given in Doyle-Connolly and Michael (2023), which covers the cases that the asymptotic distribution is normal. However, we have not discussed how an analyst, presented with a single data sample, might determine which of the two regimes applies without making too many assumptions. Doyle-Connolly and Michael (2023) show that the limit is non-normal under many common well-specified models, but relying on these examples in practice introduces parametric assumptions. A conservative approach is to form CIs under both approaches, i.e., assuming first the normal then the non-normal limit. A more refined approach requires further work.

References

Chay, K. Y. and B. E. Honore (1998). Estimation of semiparametric censored regression models: an application to changes in black-white earnings inequality during the 1960s. *Journal of Human Resources*, 4–38.

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.

Doyle-Connolly, A. and H. Michael (2023). Nonparametric estimation of the auc of an index with estimated parameters. Forthcoming; available at https://www.umass.edu/mathematics-statistics/directory/faculty/haben-michael.

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.

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.

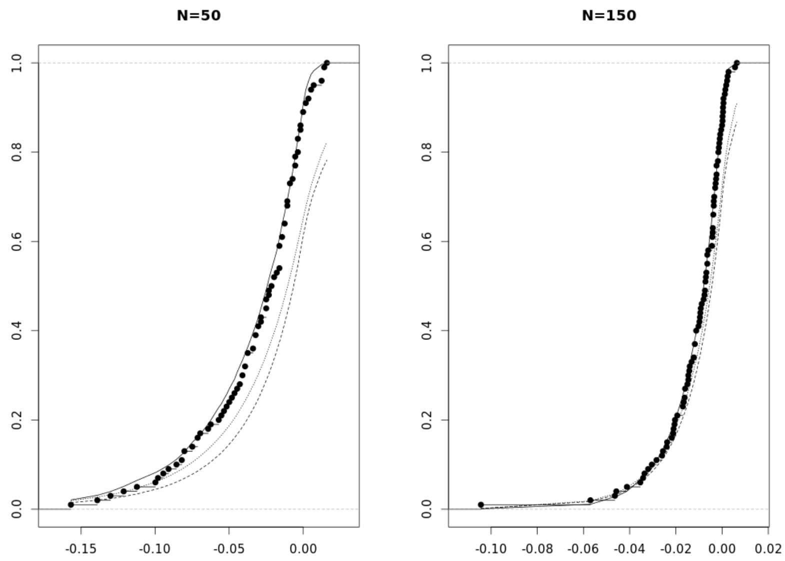
Nolan, D. and D. Pollard (1987). U-processes: rates of convergence. *The Annals of Statistics*, 780–799.

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.

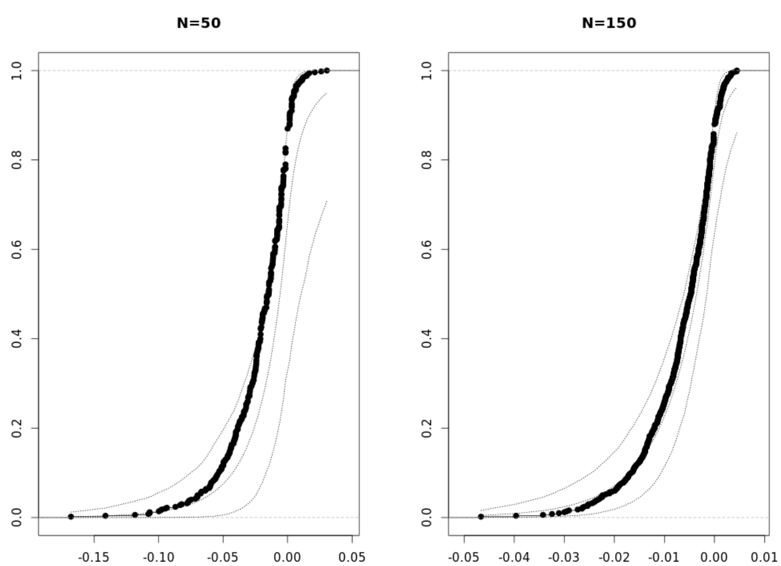
Seshan, V. E., M. Gönen, and C. B. Begg (2013). Comparing ROC curves derived from regression models. *Statistics in medicine 32*(9), 1483–1493.

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.

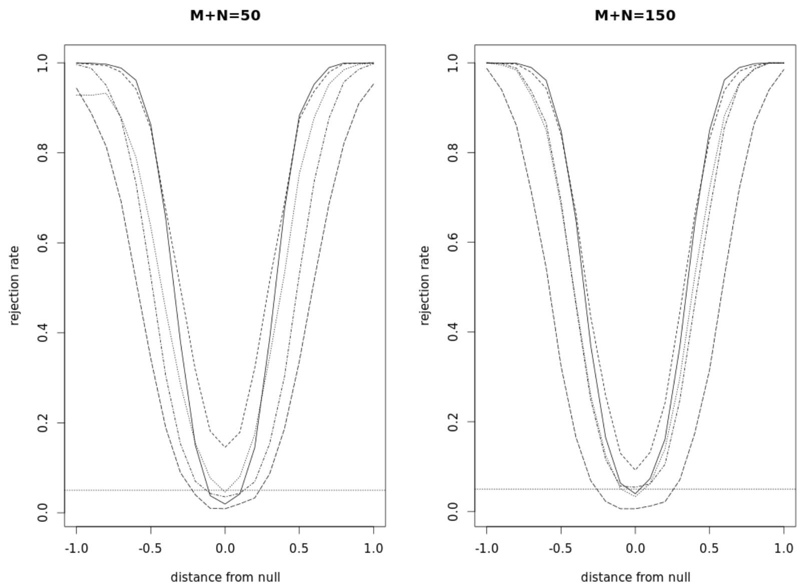


(a)

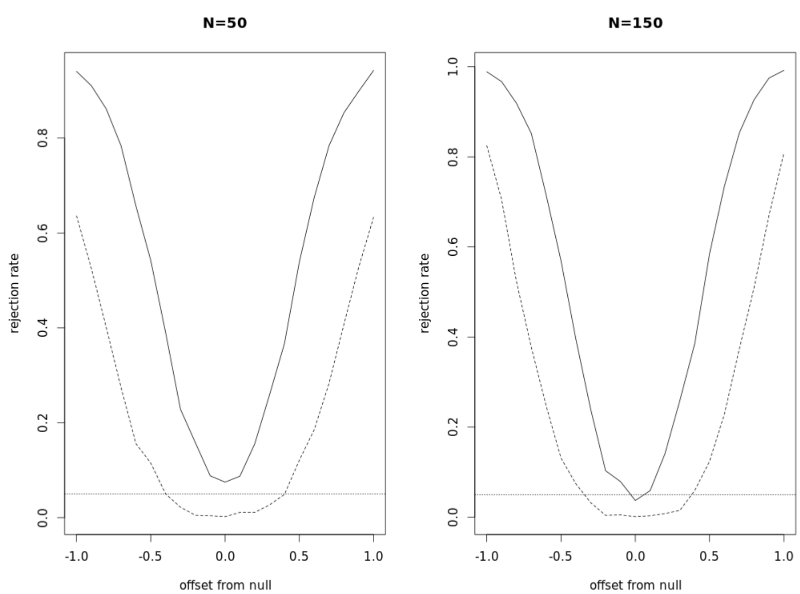


(b)

Figure 1: Observed CDF of the difference of AUCs of nested models with estimated CDFs overlaid. Shown in Fig. (1a) are the oracle CDF (solid line), parametric CDF (dashed line), and non-parametric CDF (dotted line). In (1b) the non-parametric CDF (dotted line) is given along with a 95% pointwise confidence band. The observed data is generated synthetically with the covariates modeled as normal and LDA used for the coefficient estimation. The true parameter β^* has 4 nonzero components and 2 additional components set to 0, corresponding to spurious covariates.



(a)



(b)

Figure 2: Observed rejection rate of several tests of the null $H_0 : \Delta = 0$ with conditionally gaussian data, and (2a) LDA and (2b) logistic regression for coefficient estimation. The tests in (2a) are the proposed oracle asymptotic test (solid line), parametric asymptotic test (short dashes), and non-parametric asymptotic test (dotted line), the F-test given in Demler et al. (2011) (broken line), and the standard test of DeLong et al. (1988) (long dashes). The tests in (2b) are the proposed non-parametric asymptotic test (solid line) and the standard Delong test (dashed line).