Nonparametric estimation of the auc of an index estimated from a single sample

((previously: "auc of an estimated index" but that isn't actually the target of estimation))

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

measuring the discrimination of biomarkers using the auc. ΔAUC to measure the difference in discrimination between two markers. biomarkers–index auc. often one index is based on a subset of covariates on which the other is based.

unexpected behavior when comparing ΔAUC to tests on the indexes themselves. maybe mention "bafflement" over test behavior. at the begining of 2010s.

focus here on alternative hypothesis.

useful cites: demler 2017: ΔAUC is one of the most widely used measures of discrimination.((should be "difference" in discrimination I think?)). 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

Data model: Pair (W, D), $W \in \mathbb{R}^p$ ((continuous?)), $D \in \{0, 1\}$. Denote by $X_0 \sim F, X_1 \sim G$ the RVs and distributions obtained by conditioning W on $D = 0$ and $D = 1$.

Let $((W_1, D_1), \ldots, (W_{m+n}, D_{m+n}))$ be an IID sample according to ((ref)), with the control and case variables

$$X_{01}, \ldots, X_{0m} \sim F, IID, X_{11}, \ldots, X_{1n} \sim G, IID.$$

Vectors $\hat{\beta} \in \mathbb{R}$ and $\hat{\gamma} \in \mathbb{R}$ are obtained based on the sample (the "coefficient estimation procedure"). They are assumed to have finite probability limits β^* and γ^* as $n \to \infty$. ((need to assume nonzero?))

The statistic $\widehat{\Delta \text{AUC}}$ is

$$\frac{1}{mn}\sum_{i,j}\{\hat{\beta}^T X_{0i} < \hat{\beta}^T X_{1j}\} - \frac{1}{mn}\sum_{i,j}\{\hat{\gamma}^T X_{0i} < \hat{\gamma}^T X_{1j}\}$$

An explicit probabilty model may not be specified, and often the estimation methods for $\hat{\beta}$ and $\hat{\gamma}$ imply inconsistent models. E.g., logistic models with nonzero covariates omitted from the reduced model ((ref logit example below)) 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, $\beta^{*T} x$ versus $\gamma^{*T} x$:

$$\Delta\text{AUC} = ...$$

The statistic ((ref)) may be viewed as a process of two-sample U-statistics with 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 basic U-statistics theory.

1. Under the null of no diff, $\Delta\text{AUC} = 0$, the statistic ((ref)) is often a degenerate U-statistic. The asy distribution of a degenerate U-statistic is a weighted combination of chi-squares, weights hard to estimate nonparametrically ((“literature has not presented methods for ...”)). Without this reference distribution, testing for the null directly is difficult. ((heller)) gives the asy null distribution specifically for betahat estimated by mrc method. demler for assuming coefficents are esitmated by lda with gaussian covariates. recently noted ((cite)) that asy null distribution remains intractable for common estimation methods eg logistic regression.

((pepe 2013)) provides a more uniform approach to the problem of testing for zero difference between nested index aucs. The authors suggest a more convenient often equivalent testing problem. 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 W_1, W_2 be two sets of covariates and D binary. The authors show that the null of equal AUCs of the risks,

$$P(\rho_{W_1,W_2} < ... \mid D, ...$$

holds if and only if the risk functions are equal, $\rho_{W_1,W_2} = \rho_{W_1}$. Often the coefficient estimation procedure is of secondary importance and the goal of testing the null $\Delta\text{AUC} = 0$ 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. 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 $\Delta\hat{\text{AUC}}$ were available, one might 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 ΔAUC calls for fewer modeling assumptions, improved efficiency, or something else to recommend it, may as well test risk functions.

We therefore only consider the alternative case ((ref)) 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 U-statistic is based are not IID. Non-degenerate U-statistics with estimated

parameters are typically still normal though estimation of the parameter in general affects the asy distribution. This issue is addressed in the remainder.

((maybe mention that nolan-pollard papers address both problems in almost the same setting (degree 2 ustats) but still need to nonparametrically compute the liiting chi squared distribution for the null))

3 Method

The usual approach to finding the asymptotic distribution of a U-statistic, which we adopt, is to find an asymptotically equivalent IID mean, 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,$P(\beta' X < \beta' Y)$ 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, $\Delta\hat{\mathrm{AUC}} = \theta(\hat{F}, \hat{G}, \hat{\beta}) - \theta(\hat{F}, \hat{G}, \hat{\gamma})$. We write each term as an IID sum, and later take the difference to represent $\Delta\hat{\mathrm{AUC}}$ as an IID sum. 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

$$\begin{aligned}
&\theta(\hat{F}, \hat{G}, \hat{\beta}) - \theta(F, G, \beta) \\
&= \theta(F + \delta F, G + \delta G, \beta + \delta\beta) - \theta(F, G, \beta + \delta\beta) && (1) \\
&+ \theta(F, G, \beta + \delta\beta) - \theta(F, G, \beta) && (2)
\end{aligned}$$

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

term (1): As the function $\theta(\cdot, \cdot, \beta)$ is bilinear,

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

The third and final term in the sum is $o((m + n)^{-1/2})$,

$$P(\theta(\delta F, \delta G, \beta + \delta\beta) > n^{-1/2}\epsilon) \leq P(\int d|F_n - F|(x) > n^{-1/4}\sqrt{\epsilon}) + P(\int d|G_n - G|(x) > n^{-1/4}\sqrt{\epsilon}) \to 0$$

by a DKW-type bound. ((Pf: By multivariate DKW (e.g., Serfling p.61), $P(|F_n - F|_\infty > n^{-1/4}) < c\exp(-c\sqrt{n})$))

For fixed $\beta + \delta\beta$, the first two terms are centered IID averages That the randomness in $\delta\beta$ is asymptotically negligible ((at this 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 process $\beta \to$ ... ((maybe cite pollard)).

Therefore,

$$\theta(F+\delta F, G+\delta G, \beta) - \theta(F,G,\beta) + o((m+n)^{-1/2})$$
$$= -\frac{1}{m}\sum_{i=1}^{m}(1 - G(\beta^T X_{0i}) - \theta(F,G,\beta)) + \frac{1}{n}\sum_{i=1}^{n}(F(\beta^T X_{1i}) - \theta(F,G,\beta)) + o((m+n)^{-1/2})$$

$$= \frac{1}{m+n}\sum_{i=1}^{m+n}\left(-\frac{\{D_i=0\}}{P(D=0)}(1 - G(\beta^T W_i) - \theta(F,G,\beta)) + \frac{\{D_i=1\}}{P(D=1)}(F(\beta^T W_i) - \theta(F,G,\beta))\right) + o((m+n$$

((last line doesnt maintain the $1/n$ rate? nvm it does–factor out the clsas probability estimate and end up with the product of two averages)) Known as the Hoeffding decomposition of ΔAUC, same as the first von Mises derivative. Represents term (1) as an IID sum to which the CLT may be applied to get the asymptotic distribution of ΔAUC if term (2) were negligible, e.g., if $\hat{\beta}$ were not estimated (see ((ref below) for additional scenarios when (2) is negligible)). The Delong approach in this situation is to estimate F, G using the empirical CDFs $\hat{F}, \hat{G}$, giving rise to the standard Delong statistic for inference on ΔAUC.

...

term (2): Assume $\sqrt{n}(\hat{\beta} - \beta^*) \to 0$ in probability, $\beta \mapsto \theta(F,G,\beta)$ is differentiable at β^*. Let the function $\phi_{\hat{\beta}}$ represent the estimator $\hat{\beta}$ as an IID mean

$$\hat{\beta} - \beta^* = (m+n)^{-1}\sum_{i=1}^{m+n}\phi_{\hat{\beta}}(W_i) + o((m+n)^{-1/2})$$

i.e., $\phi_{\hat{\beta}}$ is an influence function for the $\hat{\beta}$. Then (2) is

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

Putting the two parts together,

$$\theta(\hat{F},\hat{G},\hat{\beta}) - \theta(F,G,\beta)$$
$$= \frac{1}{m+n}\sum_{i=1}^{m+n}\left(-\frac{\{D_i=0\}}{P(D=0)}(1 - G(\beta^{*T} W_i) - \theta(F,G,\beta^*)) + \frac{\{D_i=1\}}{P(D=1)}(F(\beta^{*T} W_i) - \theta(F,G,\beta^*))\right)$$

$$+ \frac{\partial}{\partial\beta}\theta(F,G,\beta)\sum_{i=1}^{m+n}\phi_{\hat{\beta}}(W_i) + o_P((m+n)^{-1/2}).$$

((maybe combine into one sum))

Proposition. Assumptions: available influence function. ((continuous CDFs?)) $P(D = 0) \in (0, 1)$. non-degeneracy condition: probability limits for $\hat{\beta}$, and ... differentiability of θ at β^*, γ^*. Assertion: $(m+n)^{-1}(\theta(\hat{F}, \hat{G}, \hat{\beta}) - \theta(F, G, \beta^*))$ is asymptotically normal with mean zero and variance ((variance of a term in the IID mean)). This variance may be consistently estimated as $\sqrt{m+n}$ times the sample variance of ((ref IID mean)).

take the difference with the same representation of another estimator, $\theta(\hat{F}, \hat{G}, \hat{\gamma})$, to obtain an IID representation of $\Delta\hat{\text{AUC}}$. corollary. Assumptions: Assumptions of ((ref prop)) apply to $\hat{\beta}, \hat{\gamma}$ both, also $\beta^* \neq \gamma^*$. Then $(m+n)^{-1}(\Delta\hat{\text{AUC}} - \Delta\text{AUC})$ is asymptotically normal with mean zero and variance given by a term in the difference of IID means ((ref)). This variance may be consistently estimated as ... ((as in prop))

REMARKS

- Benefits of approximating by an iid sum:

1. As above, can de-couple the two terms of the difference ΔAUC and treat estimation the auc of an index using an estimated index.

2. $\hat{\theta}$ is an IID sum, and the sd estimate is the empirical estimator. This is itself an estimate of $\sqrt{m+n}\hat{\theta}$, not $\hat{\theta}$. So etimated parameters in ((ref single sum)) are asymptotically negligible as long as the parameter estimates are consistent and dependence is continuous. of course may affect efficiency of asy convergence. ((Just as the standard Delong estimator estimates the conditional CDFs F, G using empirical CDFs))

3. The methods apply not only to testing indexes based on nested data sets, but more generally to a comparison of any correlated AUCs with index coefficients estimated from the data, eg LDA versus logistic.

- It isn't needed that $\hat{\beta}$ be estimated by a correctly specified model, only that it has some probability limit at the parametric rate. Though procedure for obtaining the estimate $\hat{\beta}$ and the associated influence function ϕ often involve some parametric assumptions, we still term the procedure described here as "non-parametric" since the estimate ((ref)) is valid under misspecification. 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? (1) hajek parts will be the same if $\beta^* = \gamma^*$, so (1) will be $o(n^{-1/2})$. If the probability limit of $\hat{\beta}$ and $\hat{\gamma}$ are the same, then in term (2) the derivatives 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 also (1) will be $o(n^{-1/2})$. just a sufficient condition, possible that in e.g. a logit model both full and reduced are misspecified, and then the limit is normal. ((check))((also check against demler 2017 iff conditions for degeneracy))

((use of starred parameter inconsistent in this section... drop it here in favor of $\delta\beta$?))

4 Examples

4.1 No effect of coefficient estimation

In the ordinary course, the coefficient estimation can be ignored in computing the index of an auc iff the derivative ((ref)) is 0. For the difference of two AUCs

...

, the derivative of each must usually be 0 at the respective coefficient probability limits, β^*, γ^*. In that case the usual delong statistic may be used, provided of course the AUCs are distinct ((ref intro)).

4.1.1 AUC

Examples where the coefficient estimation may be ignored in estimating the AUC of an index.

Example. estimator: mrc, covariate restrictions: none/nonparametric. The maximum rank correlation method of computing the coefficients is

....

The method is non-parametric. By construction the empirical AUC is stationary at the coefficient estimates, and ((under regularity conditions)) the AUC is stationary at the probability limits, as well.

The following proposition ((highlighted by the work of Pepe)) furnishes other examples.

Proposition:

Given RVs (W, D), W continuous, D binary

1. The ROC curve of predicting D based on a real function of W is maximized pointwise over all such functions by the likelihood ratio ((...)), 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 is has the same conditional distribution given D as an increasing function of the risk of D based on W, i.e., there is a strictly increasing function $h : \mathbb{R} \to \mathbb{R}$ such that $P(\rho(x) < \xi | D = i) = P(h \circ f(W) < \xi | D = i)$ for all $\xi \in \mathbb{R}$ and $i = 0, 1$.

3. Assuming as above the derivative of the AUC is smooth at β^*, can use delong statistic if the index is related to risk by an increasing function. ((maybe introduce notation , then can say $\beta^{*T} W \sim \rho(W)$))

Proof. 1. Is the neyman pearson lemma, as pointed out by ((pepe? check if she did it first. swets.)). Let 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 $D = 1$ rejects for large values of the likelihood ratio of ((x,g)). 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 incrasing functions of the likelihood, and ((show likelihood is expit of risk)), 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, which does imply the index has the same conditional distributions as an increasing function of the risk.

Example. coefficient estimator: irrelevant, covariate restrictions: A single covariate. When there is a single covariate, $p = 1$, the β in the ((AUC formula)), 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. 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.

sub-Example. 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:

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

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

sub-Example. coefficient estimator: lda (homoskedastic), covariates: multivariate gaussian (just with gaussian data? conjecture in demler 2012 re elliptic distributions. maybe expand on this example in the misspecified lda section.) ((lda is collapsible more generally, but risk may not be an increasing function of the index without the gaussian assumption))

((maybe give proof here, or can mix it in with longer example below))

sub-Example. coefficient estimator: lda (heteroskedastic), covariates: independent exponential family data mean parameterized

Let $W_i \mid D = j$ have density $h_i(x_i)\exp(\theta_{ij}' t_i(x_i) - A_{ij})$. If the covariates are independent, the likelihood ratio is then

$$\frac{f(x \mid D = 1)}{f(x \mid D = 0)} \sim \sum_{i=1}^{n} (\theta_{i1} - \theta_{i0})^t x_i$$

((notation: n, also x's.))

If LDA is used to estimate β, then

$$\hat{\beta} \to_p ...$$

and the index at probability limit is .... If 1. the covariates have the same population variances, say $\pi_0 A_0'' + \pi_1 A_1''$, and 2. the parameter θ_{ij} is the mean A_{ij}, then

$$\beta x \sim \sum_{i=1}^{n} (A_{i1}' - A_{i0}') x_i \sim \rho(x)$$

With gaussian data as in 2 but heteroskedastic, or heteroskedastic exponential family as in 2 but not independent, the derivative of θ^* need not vanish. [[ref heteroskedastic lda example below.]]

4.1.2 difference of AUCs

For a non parametric estimator like MRC there is no difficulty. ((cite heller)) Each estimation procedure leads to a vanishing derivative.

also no problem 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. ((cite demler 2017 simulation here))

When both coefficient vectors being compared are modeled parametrically. consider specifically nested ((binary response)) models. break up into 3 cases. If neither the full model nor the reduced model is well spcified, ..., there is no reason to expect the derivative to vanish and in general coefficient estimation must be accounted for. When 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 ((ref intro)).

Finally, consider the situation that the full model is correct, reduced need not be. eg when the fuller model contains a superset of the model covariates, and the reduced model a strict subseet. This isutation is common ((give citations to simulations/data anlayses)) as in many cases correctness of the full model,

implies the reduced model usually cannot be correct

In this situation the derivative term will be nonzero and must be accounted for. For the binary response model, the requirement is that the marginalization does not break the model, equality holds in ((ref above)) ((maybe cite bridge distribution paper.)) Some examples where this requirement holds are:

- probit regession with gaussian covariates.

- gaussian lda model ((cite demler here, mahalanobis distance connection)), heteroskedastic lda model

- the logistic view of LDA: logistic regression with a gaussian mixture as covariates in the special case that beta is the lda/malanobis dist beta. This example is almost the same as the homoskedastic LDA example, since [[ref fisher lda display]] the posterior probabilities are given by ... .

convergence to 0 of the non-delong part can be slow. even on simple data like iid gaussian covariates with a null logistic model.

4.2 must account for beta estimation

Next we describe situations where must account for the β parameter estimation include: misspecification in one of the above situations ((can view ths proposed approach as adding robustness)). or, correct specified but nonzero derivative still.

Example 1. heteroskedastic gaussian lda. Suppose Gaussian linear discriminant analysis is applied to estimate beta but possibly misspecified in that the two classes may not have the same covariance. The model is

$$W|D = i \sim F_i = N_p(\mu_i, \Sigma_i)$$
$$P(D = 1) = 1 - P(D = 0) = \pi_1$$

The LDA parameter estimation procedure is to base class membership on the sign of $\beta' x$ ((ignore intercepts without loss)), where

$$\hat{\beta} = ...$$
$$\hat{\mu} = ...$$
$$\hat{\Sigma} = ...$$

the LDA parameter estimates ((ref above)), which assume a common variance for the two classes, tend in probability to

$$\beta^* = ...$$
$$\Sigma^* = ...$$

The AUC and its derivative at the starred parameters are

$$\theta(F, G, \beta^*) = ...$$
$$\theta'(F, G, \beta^*) = ...$$

The derivative is 0 iff ((eigenvector condition)). In terms of the normal means and variances, this condition is ((...)). 2 examples: 1. independent data, ((...)), 2. proportional covariance matrices. The first is already implied by the general exponential family result [[ref above]] but not the second as the observations are not independent.

show that derivative term can be unbounded. When $\Sigma \approx \sigma^*$ [[not defined yet $\Sigma = \Sigma_0 + \Sigma_1$]], the the influence function is approximately $O(|\Sigma|^{-1/2})$, the root of the inverse Fisher information, and $\theta'(F, G, \beta^*)$ is approximately $O(|\Sigma|^{1/2})$, so that the product, giving the entire non-Delong term, is approximately $O(1)$. ((maybe give quantitative bound on derivative in terms of class imbalance)) Nevertheless it is possible to push the proportion so that the entire non-Delong term $(\Sigma^*)^{-1}\theta'(F, G, \beta^*)$ has large components.

Let

$$\Sigma_0 = ...$$

Then $(\Sigma^*)^{-1}\theta'(F, G, \beta^*) \to \pm\infty\mathbb{1}$ ((write out?)) as $\pi_0 \to 1$ and $\epsilon \to 0$ simultaneously. One therefore expects that under this scenario inference based on the Delong estimator will be faulty, as verified by simulation in ((ref simulation section)). ((possible to give result so that the sign of the derivative term is controlled, so that delong can be forced either to low fpr or low power?))

5 Simulation

6 Discussion

approach extends straightforwardly to other differentiable functions of the data $f(x)$, not just the linear combination.

extends to discrete covariates (modified auc kernel)