Title: Inference on the AUC for clustered data

Abstract: The AUC is a statistic frequently used to evaluate a scalar predictor of a binary outcome. When data are correlated, Obuchowski '97 presented an estimator for the variance of the AUC statistic that remains standard in this setting. We show that this estimator is nearly the same as the leave-one-out jackknife, an alternative variance estimator based on resampling. We reduce the difference of the two variance estimators to the difference of two measures of statistical dispersion, the usual sample standard deviation versus a possibly novel statistic. This leads to a deterministic $O(1/n^2)$ bound on the difference, which is shown to be tight by an adversarial example. In as much as the jackknife is a linearization of the bootstrap, this result points to the latter as the superior method of inference for this problem.

Divide the sample into larger and smaller halves and take the difference of the means of the two halves.

1. Intro – AUC, clustering. give overview of different clustering setups in literature. examples of clustered auc applications. mention two estimators of variance. simulation showing similarity. both consistent, but not just O(1/sqrt(I)) difference. more like $O(1/I^(3.5))$, relative difference: .

   Given data consisting I iid pairs of vectors of continuous scalars

   $$X_i = (X_{i1}, \ldots, X_{im_i}), Y_i = (Y_{i1}, \ldots, Y_{in_i}), i = 1, \ldots, I$$

   The two vectors are regarded as belonging to two states, e.g., 0 and 1 or non-diseased and diseased, of ((a unit)). The quality of the vectors as predictors of the binary state is to be assessed. Common examples of correlated biomarker/response data:

   (a) repeated measurements of tumour antigens (CEA, CA15-3, TPS) as markers for progression/non-progression of breast cancer ((ref emir 2000))

   (b) two measurements of the distortion product otoacoustic emissions taken from the left and right ears ((think this should be ears)) of each patient, response: neonatal hearing impairment ((wu 2019))

   (c) repeated measurements of levels of vascular enothelial growth factor and a soluble fragment of Cytokeratin 19 as prognostic factors for progression of non-small cell lung cancer ((ref wu wang 2011))

   One way to assess the quality of the predictors is the average probability that a non-diseased observation is less than a diseased observation, where the average is taken over comparisons of elements of a non-diseased cluster against those of an independent diseased cluster((should be: the mean AUC between the controls of one cluster and the cases of an independent cluster))

   $$\theta = E\left(\frac{1}{m_i n_k}\sum_{j=1}^{m_i}\sum_{l=1}^{n_k}\{X_{ij} < Y_{kl}\}\right).$$

((need to treat mi,nk. are these random? or make them fixed)). Definition ((ref above)) is one way of generalizing the AUC, a way to evaluate a scalar predictor of a binary outcome, to clustered observations. When $m_i = n_i = 1, i = 1, \ldots, I$ in ((ref above)), the AUC may be defined as $P(X_i < Y_j), i \neq j$, the probability that a diseased observation exceeds an independent non-diseased observation. When the observations in a cluster have the same marginal distribution, the two definitions agree.

The estimate of the AUC presented in ((refs)) is

$$\overline{\phi}_{\cdot\cdot} = I^{-2} \sum_{i,j} \phi_{ij}$$

where

$$\phi_{ij} = \frac{1}{mn} \sum_{k,l} \{x_{ik} < y_{jl}\}, 1 \leq i, j \leq I$$

The estimator has an $O(1/I)$ bias due to the intra-cluster terms ϕ_{ii}. ((these terms could be omitted from the estimator, though we follow the original definition in our analysis))

	case	control
patient # 1	X_1	
$\vdots$	$\vdots$	
patient # k	X_k	
patient # k+1		Y_{k+1}
$\vdots$		$\vdots$
patient # I		Y_I

Lee et al.

	case	control
patient # 1	$(X_{11}, \ldots, X_{1m_1})$	
$\vdots$	$\vdots$	$\vdots$
patient # k	$(X_{k1}, \ldots, X_{km_I})$	
patient # k+1		$(Y_{(k+1)1}, \ldots, Y_{(k+1)n_{k+}}$
$\vdots$		$\vdots$
patient # I		$(Y_{I1}, \ldots, Y_{In_I})$

Obuchowski:

	case	control
patient # 1	$(X_{11}, \ldots, X_{1m_1})$	$(Y_{11}, \ldots, Y_{1n_1})$
$\vdots$	$\vdots$	$\vdots$
patient # I	$(X_{I1}, \ldots, X_{Im_I})$	$(Y_{I1}, \ldots, Y_{In_I})$

((obu ref)) presents an estimator of the variance of ((ref pop auc)) enabling inferences to be drawn. Alternative estimators of the variance are resampling methods such as the bootstrap and the jackknife. Fig (()) presents a comparison of ((obu estimator)) and jackknife estimator using synthetic multivariate normal data. The two estimates are nearly identical. Fig (()) suggests that the difference of the two variance estimates is somewhere between $O(1/I^2)$ and $O(1/I^{2.5})$.

Below we establish that the difference between the two variance estiamtors is $O(1/I^2)$. We also give an adversarial example achieving this bound. Inasmuch as the jackknife may be viewed as a linearization of the bootstrap ((ref)), the result points to the bootstrap for carrying out inference in this setting ((ref model above)).

2. Body

   (a) obu estimator and jk, give objective to maximize

   Let

$$
\begin{aligned}
x \in \mathbb{R}^{I\times m}, y \in \mathbb{R}^{I\times n}\\
\phi_{ij} = \frac{1}{mn}\sum_{k,l}\{x_{ik} < y_{jl}\}
\end{aligned}
$$

Were non-diseased cluster x_i and diseased cluster y_i independent, the estimator would be a two-sample U-statistic. The variance estimator proposed in ((obu ref)) extends the usual U-statistic variance estimator to account for this dependence.

The normalized variance of $\hat{\theta}$ is

$$
Var(\sqrt{I}\hat{\theta}) = \frac{1}{I^3}\sum_{i,j,k,l} Cov(\phi_{ij}, \phi_{kl})
$$

The sum consists of $O(I^4)$ terms in which i, j, k, l are all distinct, which are 0 since distinct clusters are independent; $O(I^3)$ terms for which $|\{i, j, k, l\}| = 3$; and an asymptotically negligible number of remaining terms.

Among those terms for which $|\{i, j, k, l\}| = 3$ are, first, those of the form

$$
\begin{aligned}
Cov(\phi_{ij}, \phi_{ik}) &= E(\phi(X_i, Y_j)\phi(X_i, Y_k)) - \hat{\theta}^2\\
&= E(f_{10}(X_i)^2) - \hat{\theta}^2,
\end{aligned}
$$

((notation $\phi(x, y)$)) writing $f_{10}(X_i) = E\phi(x, Y_j)|_{x=X_i}$. ((Sen reference)) shows that $\frac{1}{I}\sum_j \phi(X_i, Y_j) \to_p f_{10}(X_i)$, leading to $\frac{1}{I^2}\sum_i(\overline{\phi}_{i\cdot} - \hat{\theta})^2$ as a consistent estimator of the covariance ((ref above display)).

A second type for which $|\{i, j, k, l\}| = 3$ is $Cov(\phi_{ij}, \phi_{kj})$, which may similarly be estimated by $\frac{1}{I^2}\sum_i(\overline{\phi}_{\cdot j} - \hat{\theta})^2$. Finally, to account for covariances of the form $Cov(\phi_{ij}, \phi_{ki}) = Cov(\phi_{ij}, \phi_{ji})$, representing the intra-cluster dependence, ((obu ref)) adds the analogous term $\frac{2}{I}\sum_i(\overline{\phi}_{i\cdot} - \hat{\theta})(\overline{\phi}_{\cdot i} - \hat{\theta})$. The final variance estimator presented in ((obuchowski ref)) is:

$$
\hat{\sigma}^2_{obu} = \frac{1}{I(I-1)}\sum_{j=1}^{I}(\overline{\phi}_{j\cdot} + \overline{\phi}_{\cdot j} - 2\hat{\theta})^2.
$$

An additional contribution of ((obu ref)) is to account for variable cluster sizes, though the focus here is on fixed cluster sizes.

several estimates of the variance, depending on the cluster/correlation. these include... more recent examples include ...

The jackknife is a general-purpose technique for estimating the variance of a statistic. For a statistic based on an i.i.d. sample of size I, in this application the observations are the cluster pairs $(x_1, y_1), \ldots, (x_I, y_I)$, I "leave-one-out" statistics are formed,

$$\overline{\phi}_{\cdot\cdot,-j} = ..., j = 1, \ldots, I$$

"Pseudo-observations" are obtained as ..., and the jackknife variance estimate is the unbiased sample variance of these pseudo-observations

$$\hat{\sigma}^2_{jk} = ...$$

in terms of ϕ

$$\hat{\sigma}^2_{jk} = \frac{I}{(I-1)^3} \sum_{j=1}^{I} \left(\overline{\phi}_{j\cdot} + \overline{\phi}_{\cdot j} - 2\hat{\theta} - \frac{1}{I} \left(\phi_{jj} - \frac{\mathrm{tr}(\phi)}{I} \right) \right)^2$$

The difference is

$$\frac{2I-1}{I(I-1)^3} \sum_j (\overline{\phi}_{j\cdot} + \overline{\phi}_{\cdot j} - 2\hat{\theta})^2 - \frac{2}{(I-1)^3} \sum_j \left(\overline{\phi}_{j\cdot} + \overline{\phi}_{\cdot j} - 2\hat{\theta} \right) \left(\phi_{jj} - \frac{\mathrm{tr}(\phi)}{I} \right) + \frac{1}{I(I-1)^3} \sum_j \left(\phi_{jj} -$$

The difference appears to be $O(1/I^2)$, the relative difference $O(1/I^{3/2})$. (Q is a difference of 2 projections...do these correspond to jk estimator and the obu estimator?? dont think so Q is actually difference of a sample variance and a covariance)

(b) description of phi.

general unbalanced case

balanced case, m and n fixed, then ϕ belongs to the set of matrices

$$\frac{1}{mn} \underset{I \times In}{V} \underset{In \times I}{P} : 0 \leq v_1 \leq v_2 \leq \ldots v_{In} \leq m, \sum_j P_{ij} = 1, \sum_i P_{ij} = n, i = 1, \ldots, I$$

i.e., letting U represent the upper right $I \times I$ matrix,

$$\begin{aligned} \frac{1}{mn} \underset{I \times In}{A} \underset{I \times I}{U} \underset{Im \times I}{B} &: A_{ij}, B_{ij} \in \{0, 1\} \\ & A_{i\cdot} = n, A_{\cdot j} = 1 \\ & B_{i\cdot} = 1, B_{\cdot j} = m, 1 \leq i, j \leq I \end{aligned}$$

Letting $\mathbb{P}_n(f(Y_k)) = \frac{1}{n}\sum_{l=1}^{n} f(y_{kl})$ denote expectation with respect to the empirical distribution of the diseased clusters $y_1, \ldots, y_I$, $\hat{F}_i$ be the empirical CDF of cluster x_i

$$\phi_{ij} = \mathbb{P}_n \hat{F}_i(Y_j)$$

((fig of PF decomposition))

(c) quad form pictures of summands in 3x3 case, kronecker and delta descriptions. symmetric-looking representation of an elt, some consequences of the symmetry, how these make sense intuitively.

block structure. View Q is an $I \times I$ matrix of $I \times I$ blocks. viewing qf as a sum of qfs.

Theorem 1. *i. elt description of Q shows symmetric in p, q, r, s, where p, q indexes the block and r, s indexes within the block. The number of pairs or triples from (p, q, r, s) that are all equal is invariant under any permutations π on $1, \ldots, I$, so $Q_{pqrs} = Q_{\pi(p)\pi(q)\pi(r)\pi(s)}$, and*

$$\phi^t Q \phi = \sum_{p,q,r,s} Q_{pqrs}\phi_{pr}\phi_{qs} = \sum_{p,q,r,s} Q_{\pi(p)\pi(q)\pi(r)\pi(s)}\phi_{\pi(p)\pi(r)}\phi_{\pi(q)\pi(s)}$$

$$= \sum_{p,q,r,s} Q_{pqrs}\phi_{\pi(p)\pi(r)}\phi_{\pi(q)\pi(s)}$$

ii. corollay: blocks are symmetric. this is switching r, s in Q_{pqrs}. similarly for symmetry about anti-diagonal.

iii. Q invariant on following symmetries. any permutation applied in common to the rows and columns of the input vector, viewed as a matrix. this corresponds to rearranging the iid input clusters.

iv. Q invariant on 1) transposition. proof from $Q[p, q, r, s]$ symmetry: switch p and q, and switch r and s. corresponds to switching roles of x and y. 2) reflection in the anti-diagonal. pf:? interpretation:?.Along with the identity transformation, and 180 degree rotation (which is a permutation symmetry), these form a 4-member subgroup of S4, symmetry group of the square.

v. blocks sum to 0 proven in notes using kronecker notation

theorem on operator

i. higher powers of Q, (Q.lambda)2 is an orthog proj matrix. nullspace, rank, characteristic eqn. Q is difference of two projection matrices. rank of Q very small but need to understand phi to bound phi.Q.phi.

ii. For $k \geq 0$, $Q^{2k+1} = \lambda^{2k} Q$ and $Q^{2k} = \lambda^{2(k-1)} Q^2$ corollary: $(Q/\lambda)^2$ is an orthogonal projection onto the column space of Q^2, which is the same as the column space of Q. corollary: Q/λ

is a differene of two projection matrices onto two ((dimensions)) orthogonal subsapces.

iii. for $k > 0$

$$\mathrm{tr}(Q^{2k}) = \begin{cases} 2(I-1)\lambda^{2k}, & k \neq 0, I \neq 2 \\ 0, & k \neq 0, I = 2 \\ I, & k = 0 \end{cases}$$

iv. The null space of Q consists of

   A. ...

   B. ...

   C. ...

v. 2*(I-1) nonzero eigenvalues, all equal in magnitude to $\lambda = \sqrt{\frac{I-2}{2I^2}}$, half positive and half neg.

vi. eigenvectors for nonzero evals sum to 0, since row sums of Q sum to 0.

vii. The characteristic polynomial of Q is

$$\frac{(-1)^I}{2^{I-1}} x^{I^2-2(I-1)}(2x^2 - I^2(I-2))^{I-1}.$$

Let column j of WW have a 1 in position $(jmodI) * I + floor(j/I)$, and 0 elsewhere.(check in R). Then the symmetry of $Q[p, q, r, s]$ in q, s implies WW is a right identity for Q, and symmetry in p, r, and $WW = WW^t$, implies it is a left identity. So $WW - E$ is in the nullspace of Q. $WW - E$ has $I(I-1)$ nonzero columns and for each column its negation is also present, leaving $I(I-1)/2$-dimensional subsapce of the nullspace of Q in $WW - E$.

isotropic vectors and ortho decomposition. row-constant matrices and row arithmetic progressions, bases.

remark re stirling numbers

corollary of theorem on symmetries

Corollary 1. *2I phi.vecs kronecker(E,ones),kronecker(ones,E) mutually orthogonal in Q basis. ((also, diagonal))((also, upper right triangular))*

follows from blocks summing to 0 and Q(p,q,r,s) symmetry in arguments. shows row and col sums of Q are 0. $(e_j \otimes \mathbb{1})^t Q(e_k \otimes \mathbb{1}) = 0$ is the same as the j, k block of Q summing to 0. This is fixing p, q as j, k and summing over r, s in the Q elt notation. by Q argument symmetry, we can replace the role of p with r and q with s shows $(\mathbb{1} \otimes e_j)^t Q(\mathbb{1} \otimes e_k) = 0$. Instead switching the roles of q and s shows $(e_j \otimes \mathbb{1})^t Q(\mathbb{1} \otimes e_k) = 0$.

(These are Q-orthogonal like B and C (actually first I of these are B)–actually these give C: it is a linear combination of vectors with constant columns. so I should have been projecting onto this basis.)

$O(1/I^2) counterexample$

3. Conclude/Discussion

Proof. Let $c_j, j \geq 1$, as the coefficients of x^{I^2-j} in the characteristic equation of Q, with $c_0 = 1$. The Newton Identities expressing c_k in terms of the traces of powers of Q are ((ref V.V. Prasolov, Problems and Theorems in Linear Algebra (American Mathematical Society, Providence, RI, 1994) sec 4.1))

$$c_k = \sum_{j=1}^{k} (-1)^{I^2-j} \sum_{\substack{i_1+\ldots+i_j=k \\ i_1\geq 1,\ldots,i_j\geq 1}} \prod_{l=1}^{j} \frac{\mathrm{tr}(Q^l)}{l}.$$

Since $\mathrm{tr}(Q^{2k+1}) = 0, k \geq 0$, each sum in ((ref newton identities)) is 0 when k is odd. Otherwise,

$$\begin{aligned}
(-1)^{I^2} c_{2k} &= \sum_{j=1}^{k} (-1)^j \sum_{\substack{i_1+\ldots+i_j=2k \\ i_1\geq 1,\ldots,i_j\geq 1}} \prod_{l=1}^{j} \frac{\mathrm{tr}(Q^l)}{l} \\
&= \sum_{j=1}^{k} (-1)^j \sum_{\substack{i_1+\ldots+i_j=k \\ i_1\geq 1,\ldots,i_j\geq 1}} \prod_{l=1}^{j} \frac{\mathrm{tr}(Q^{2l})}{2l} \\
&= \left(\frac{I-2}{2I^2}\right)^k \sum_{j=1}^{k} (-1)^j \frac{(I-1)^j}{j!} \sum_{\substack{i_1+\ldots+i_j=k \\ i_1\geq 1,\ldots,i_j\geq 1}} \frac{1}{i_1 \cdot \ldots \cdot i_j} \\
&= \left(\frac{I-2}{2I^2}\right)^k \sum_{j=1}^{k} (-1)^j \frac{(I-1)^j}{k!} |s_j^k| \\
&= \left(\frac{I-2}{2I^2}\right)^k (-1)^{I+k} \binom{I-1}{k},
\end{aligned}$$

((use k or j consistently for index consistently)) denoting by $|s_j^k|$ the unsigned Stirling number of the first kind ((ref for identity)).

Therefore the characteristic equation is

$$\begin{aligned}
\sum_{j=0}^{I^2} c_j x^{I^2-j} &= \sum_{j=0}^{I-1} c_{2j} x^{I^2-2j} \\
&= \sum_{j=0}^{I-1} \left(\frac{I-2}{2I^2}\right)^j (-1)^{I+j} \binom{I-1}{j} x^{I^2-2j} \\
&= (-1)^I x^{I^2-2(I-1)} \sum_{j=0}^{I-1} \binom{I-1}{j} x^{2(I-1-j)} \left(-\frac{I-2}{2I^2}\right)^j \\
&= (-1)^I x^{I^2-2(I-1)} \left(x^2 - \frac{I-2}{2I^2}\right)^{I-1}.
\end{aligned}$$

□