Testing for a difference in AUCs based on LDA fitted values

ABSTRACT: When the AUCs of summary biomarkers are compared, the data used for inference is often clustered. For example, patients may contribute a series of measurements from several hospital visits, in which case the measurements belonging to a given patient may be strongly auto-correlated. However, current tests for a difference in summary biomarkers assume IID data. We extend the test for a difference in summary biomarkers to allow for dependence among the observations. Our "cluster-robust" test allows for both dependence among a given subject's measurements and also across arms, such as when a given subject changes disease status and contributes measurements both when diseased and non-diseased. While assuming that the marginal distribution of the observations is elliptically contoured, we make no strong assumptions on the nature of the dependence. Our cluster-robust test therefore prevous tests, which are either restricted to IID data, or to AUCs based on simple markers and not summary biomarkers.

KEYWORDS: AUC, Clustered data, Elliptically contoured distributions

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 construction of scientists, being constructed from other markers by a formula. Very often the marker takes the form $\hat{\beta}^T x$, where x is a vector of subject characteristics and $\hat{\beta}$ is obtained by fitting a binary classification model to case and control data. Comparison of markers then takes the form of comparing two sets of patient characteristics x and y, with corresponding fitted values $\hat{\beta}^T x, \hat{\gamma}^T y$. The characteristics are typically nested, $x \subset y$, as when investigating the impact on discrimination of the additional factors that lie in y but not 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).

Inference for the difference in AUCs, and in particular testing the null of no difference, is relatively complicated (Michael et al., 2024). The asymptotic distribution of the empirical test statistic is in general non-normal. An alternative approach is to bring parametric modeling to bear and use a simpler test statatistic. Extending Su and Liu (1993), Doyle-Connolley and Michael (2024) shows that when the data is multivariate Gaussian, the optimal decision rule, linear or otherwise, is the LDA decision rule. Demler et al. (2011) then shows that testing for a difference in the AUCs of the resulting fitted values can be replaced with testing for a difference of Mahalanobis distances, a well understood classical problem. Combining the optimality property and the equivalence property shows that it makes sense to look at the Mahalanobis distance to compare discrimination whenever data can be assumed Gaussian.

We extend the result of Demler et al. (2011) in two ways. First we show that the difference in LDA fit AUCs is equivalent to the difference in Mahalanobis distances not just for Gaussian data but many elliptically contoured random vectors. Then we propose a test of the null when the data is clustered.

In Section 2 we carry out the extensions of the Demler et al. (2011) result, proposing a hypothesis test for the difference in LDA fit AUCs that is valid for possibly clustered elliptically contoured data. In Section 3 we examine the test's finite-sample coverage and power by simulation relative to competing methods. . Software implementing the proposed test as well as the routines for the simulations in Section 3 are publicly available at the website of the first author.

2 Theory

Suppose an analyst wishes to test for the improvement in discrimination between controls and cases, if any, offered by using a new marker over a currently used marker or set of markers. The analyst must first decide on the criterion for discrimination. Let X denote the marker or markers when drawn from a control population, and Y when drawn independently from a case population. If X and Y are scalar, $X, Y \in \mathbb{R}$, then a standard summary of the marker's capacity to discriminate between the controls and cases is the AUC,

$$P(X < Y) + \frac{1}{2}P(X = Y). \tag{1}$$

When, as is often the case, X and Y are vectors, the AUC is often computed on a summary statistic formed by combining their components. If X and Y are multivariate Gaussian with a common variance,

$$X \sim N_k(\mu_x, \Sigma), Y \sim N_k(\mu_y, \Sigma) \tag{2}$$

the optimal way to combine the components (Su and Liu, 1993; Doyle-Connolley and Michael, 2024) is the LDA fitted values ${\beta_{LDA}}^T X, {\beta_{LDA}}^T Y$, where

$$\beta_{LDA} = \Sigma^{-1}(\mu_y - \mu_x) \tag{3}$$

This summary marker is optimal in the sense that any other real function f of the marker offers less discrimination, as assessed by the AUC, than than the LDA fitted values,

$$|P(f(X) < f(Y)) - \frac{1}{2}| < |P({\beta_{LDA}}^T X < {\beta_{LDA}}^T Y) - \frac{1}{2}|.$$

Therefore, a straightforward way for the analyst to test for the improvement in discrimination when the data is Gaussian is to compare the AUCs of the LDA fits. Let the combined markers be written as $X = (X_1, X_2), Y = (Y_1, Y_2)$, where X_1 and Y_1 represent the old markers when drawn from a control and case population, respectively, and X_2 and Y_2 the new markers that are under scrutiny. Let $\beta_{LDA1} = \Sigma_{11}^{-1}(\mu_{y1} - \mu_{x1})$ be the LDA coefficients of the reduced data X_1 and Y_1. The difference in LDA AUCs is

$$\Delta AUC = P({\beta_{LDA}}^T X < {\beta_{LDA}}^T Y) - P({\beta_{LDA}}_1^T X_1 < {\beta_{LDA}}_1^T Y_1) \tag{4}$$

and the analyst is interested in testing $H_0 : \Delta AUC = 0$.

Implementing this test using off-the-shelf methods has led to faulty inferences. Suppose the analyst has a set of control and case markers known or assumed to be homoscedastic Gaussian,

$$\begin{aligned} X_1, \ldots, X_{n_x} &\sim N(\mu_x, \Sigma) \\ Y_1, \ldots, Y_{n_x} &\sim N(\mu_y, \Sigma) \end{aligned} \tag{5}$$

Often the analyst uses this same data to estimate the LDA coefficients (3) and resulting fits,

$$\begin{aligned} &\hat{\beta}_{LDA}^T X_1, \ldots, \hat{\beta}_{LDA}^T X_{n_x}, \hat{\beta}_{LDA}^T Y_1, \ldots, \hat{\beta}_{LDA}^T Y_{n_y}, \\ &\text{where } \hat{\beta}_{LDA} = \hat{\Sigma}^{-1}(\hat{\mu}_y - \hat{\mu}_x), \end{aligned} \tag{6}$$

as to test the null $H_0 : \Delta AUC = 0$. The standard test of the null (DeLong et al., 1988), however, assumes independent observations, and this assumption is violated by the common $\hat{\beta}_{LDA}$ in (6). The consequences of this violation, including loss of power, are not specific to the Gaussian or LDA setups considered here, and have been extensively discussed in the biostatistics literature, see, e.g., Seshan et al. (2013); Tzoulaki et al. (2009); Demler et al. (2012); Heller et al. (2017) and the references there.

As a remedy, Demler et al. (2011) bypass the standard test of DeLong et al. (1988). Since the difference in LDA AUCs for the data (5) is simply

$$\begin{aligned} &P({\beta_{LDA}}^T (Y - X) > 0) - P({\beta_{LDA}}_1^T (Y_1 - X_1) > 0) \\ &= \Phi((\mu_Y - \mu_X)^T \Sigma^{-1} (\mu_Y - \mu_X)/\sqrt{2}) - \Phi((\mu_{Y1} - \mu_{X1})^T (\Sigma_{11})^{-1} (\mu_{Y1} - \mu_{X1})/\sqrt{2}) \\ &= \Phi(\Delta/\sqrt{2}) - \Phi(\Delta_1/\sqrt{2}), \end{aligned} \tag{7}$$

testing the null $H_0 : AUC = AUC_1$ is the same as testing $\Delta = \Delta_1$, where $\Delta = (\mu_Y - \mu_X)^T \Sigma^{-1}(\mu_Y - \mu_X)$ and $\Delta_1 = (\mu_{Y1} - \mu_{X1})^T (\Sigma_{11})^{-1}(\mu_{Y1} - \mu_{X1})$ are the Mahalanobis distances between the control and case distributions for the full and initial data. We formulate this equivalence more explicitly, using the fact that the quantities in (7) only depend on the distribution of the difference $W = Y - X \sim N(\mu_Y - \mu_X, 2\Sigma)$:

Given a family of random vectors indexed by $\theta \in \Theta$, let $\mu(\theta) = E_\theta(X)$, $\Sigma(\theta) = Var_\theta(X)$, and $\beta_{LDA}(\theta) = \Sigma(\theta)^{-1}\mu(\theta)$. We say that the Equivalence Principle holds for the family when, for all $\theta \in \Theta$, the LDA AUC difference parameter vanishes,

$$P_\theta(\beta_{LDA}(\theta)^T W > 0) - P_\theta(\beta_{LDA1}(\theta)^T W_1 > 0) \tag{8}$$

$$= P_\theta(\mu(\theta)^T \Sigma(\theta)^{-1} W > 0) - P_\theta(\mu(\theta)_1^T \Sigma(\theta)_{11}^{-1} W_1 > 0) = 0, \tag{9}$$

if and only if the Mahalanobis distance difference parameter vanishes,

$$\mu(\theta)^T \Sigma(\theta)^{-1}\mu(\theta) - \mu(\theta)_1^T \Sigma(\theta)_{11}^{-1}\mu(\theta)_1 = 0. \tag{10}$$

Demler et al. (2011) show that the Equivalence Principle holds for multivariate Gaussians.

In the next section we extend the Equivalence Principleto elliptically contoured distributions. That is, we show that when the data are drawn from an elliptically contoured distribution with second moments, testing for a difference in LDA AUCs is the same as testing for a difference in Mahalanobis distance. To carry out the latter test in this more general setting, we can no longer rely the F-test Demler et al. propose for Gaussian data. In the subsequent section we give an asymptotic linearization of the difference in Mahalanobis distance that is valid for non-Gaussian data. Moreover, using cluster-robust CLTs we can extend the use of our test to dependent data, as we do in Section 2.3.

2.1 Elliptical distributions

A p-dimensional random vector X with variance $\Sigma > 0$ is said to be ellipitcally distributed when it has the representation (Cambanis et al., 1981)

$$X \sim \mu + R\sqrt{c}\Sigma^{1/2}U, \tag{11}$$

where $\mu = E(X)$, R is a nonnegative random variable, $\Sigma^{1/2}$ is the postive definite square root of Σ, U is a random vector independent of R that is uniformly distributed on the sphere in $\mathbb{R}^p$, and $c = p/ER^2$. radial variable r determines the characteristics of the distribution beyond the first 2 moments, e.g., normal, t, the tail behavior. An elliptically distributed random vector is a spherically symmetric random vector after application of an affine transformation.

In Demler 2011 the authors conjecture that the equivalence of maha distance and the lda auc for normal random vectors may hold more generally for ellipitcally distributed random vectors. In general this conjecture does not hold. Suppose that the radial function r is bounded, say uniform on $[0, 1]$, and for simplicity take the covariance Σ to be the identity. Let X, Y be samples from this family differing in their means μ_y, μ_x. The distributions look like unit spheres at the tips of μ_Y, μ_X. When μ_Y and μ_X are far apart relative to the unit radius spheres, $Y - X \approx \mu_Y - \mu_X$, and so ${\beta_{LDA}}^T(Y - X) \approx |\mu_Y - \mu_X|^2$ has no chance of being negative. Consequently both terms of the LDA AUC difference parameter (8) are 1 and their difference 0, while nothing at all prevents the Mahalanobis difference from being > 0 or even large. The problem is that, unlike the Mahalanobis distance, the AUC as a measure of discrimination maxes out with data that is perfectly separated. In such a situation testing for no Mahalanobis difference as a substitute for a test of no AUC LDA difference would lead to Type 1 errors in the original testing problem.

However, ruling out perfectly separable distributions, the conjecture does hold, as well as the converse.

Theorem 1. *Let Θ represent an ellipitcal location-scale family, i.e., a family of ellipitcal distributions closed under affine transformations. Assume that the radial function R has nonzero density on all of $[0, \infty)$. Then the Equivalence Principle holds. Conversely, if the Equivalence Principle holds for a family of distributions closed under affine transformations, then the distributions in the family are elliptical.*

Proof. The LDA AUC difference parameter is

$$
\begin{aligned}
&P_{\mu,\Sigma}({\beta_{LDA}}^T X > 0) - P_{\mu,\Sigma}({\beta_{LDA}}_1^T X_1 > 0) && (12)\\
&= P(\mu^T\Sigma^{-1}(\mu + R\sqrt{c}\Sigma^{1/2}) > 0) - P(\mu_1^T\Sigma_{11}^{-1}(\mu_1 + R\sqrt{c}(\Sigma^{1/2})_{1,1:2}) > 0) && (13)\\
&= P(\Delta + R\sqrt{c\Delta}\frac{\mu^T\Sigma^{-1/2}}{|\mu^T\Sigma^{-1/2}|}U > 0) - P(\Delta_1 + R\sqrt{c\Delta_1}\frac{\mu_1^T\Sigma_{11}^{-1}(\Sigma^{-1/2})_{1,1:2}}{|\mu_1^T\Sigma_{11}^{-1}(\Sigma^{-1/2})_{1,1:2}|}U > 0), && (14)
\end{aligned}
$$

with the last line using $|\mu^T\Sigma^{-1/2}|^2 = \Delta$ and $|\mu_1^T\Sigma_{11}^{-1}(\Sigma^{-1/2})_{1,1:2}|^2 = \Delta_1$. Since U is spherically symmetric, its projection onto any given unit vector in $\mathbb{R}^p$ has the same distribution as onto any other. Let ΠU denote an RV with this common distribution. Assume first that $\Delta \neq 0, \Delta_1 \neq 0$. The last line is

$$
\begin{aligned}
&P(\Delta + R\sqrt{c\Delta}\Pi U > 0) - P(\Delta_1 + R\sqrt{c\Delta_1}\Pi U > 0) && (15)\\
&= P(R\sqrt{c}\Pi U > -\sqrt{\Delta}) - P(R\sqrt{c}\Pi U > -\sqrt{\Delta_1}). && (16)
\end{aligned}
$$

Since it is assumed that the support of R is $[0, \infty)$, the CDF of $R\sqrt{c}\Pi U$ is everywhere strictly increasing. Therefore the LDA AUC difference (16) vanishes if and only if the Mahalanobis difference $\Delta - \Delta_1$ vanishes. For those θ where $\Delta = 0$, also $\Delta_1 = 0$ since $\Sigma > 0$, and so the LDA AUC difference (16) is also 0. Similarly, where $\Delta \neq 0$ but $\Delta_1 = 0$, the LDA AUC difference is $P(R\sqrt{c}\Pi U > -\sqrt{\Delta}) - 1/2 \neq 0$.

Conversely, suppose the Equivalence Principle holds for the family Θ. Then for any RV X in the family with mean μ and variance Σ, $\Delta = \Delta_1$ implies

$$
\begin{aligned}
0 &= P_{\mu,\Sigma}(\mu^T\Sigma^{-1}X > 0) - P_{\mu,\Sigma}(\mu_1^T\Sigma_{11}^{-1}X_1 > 0) && (17)\\
&= P(\mu^T\Sigma^{-1/2}Z > -\Delta) - P(\mu_1^T\Sigma_{11}^{-1/2}Z_1 > -\Delta_1), && (18)
\end{aligned}
$$

where $Z = \Sigma^{-1/2}(X-\mu)$ and Z_1 is the first k components. Taking the mean to be $c\mu$ for fixed μ and $c \in \mathbb{R}, c \neq 0$, the equality of the resulting maha distances is unaffected, $c^2\Delta = c^2\Delta_1$ iff $\Delta = \Delta_1$. If $c > 0$, by the above $\Delta = \Delta_1$ implies $P(\mu^T\Sigma^{-1/2}Z > -c\Delta) - P(\mu_1^T\Sigma_{11}^{-1/2}Z_1 > -c\Delta_1)$, so the CDFs of $\mu^T\Sigma^{-1/2}Z$ and $\mu_1^T\Sigma_{11}^{-1/2}Z_1$ agree on all negative quantiles. If $c < 0$, $\Delta = \Delta_1$ implies $P(\mu^T\Sigma^{-1/2}Z < -c\Delta) - P(\mu_1^T\Sigma_{11}^{-1/2}Z_1 < -c\Delta_1)$, so the CDFs of $\mu^T\Sigma^{-1/2}Z$ and $\mu_1^T\Sigma_{11}^{-1/2}Z_1$ agree on all positive quantiles. Therefore, for any $\mu, \Sigma > 0$,

$$
\Delta = \Delta_1 \text{ implies } \mu^T\Sigma^{-1/2}Z \overset{d}{=} \mu_1^T\Sigma_{11}^{-1/2}Z_1. \tag{19}
$$

From (19), we deduce that Z is spherical in two stages. First we show the distribution of the projection of Z_1 onto a direction in $\mathbb{R}^k$ does not depend on the direction, so that Z_1 is spherical. Then we show that the projections of Z onto any direction in $\mathbb{R}^p$ have the same distribution as the common distribution of the projections of Z_1.

1. Let $v \in S^{k-1}$ be a fixed direction, and let $u_1 \in S^{k-1}$ be a direction in the same proper half-space as v, i.e., $(u_1, v) > 0$. This half-space is the same as the set $\{Av : A > 0\}$, so let $k \times k$ $A > 0$ satisfy $Au_1 = v$. With $u = (u_1, u_2) \in S^{p-1}$ as before, $u_1 \in S^{k-1}$ implies $u_2 = 0 \in \mathbb{R}^{p-k}$. Let

$$
\Sigma = \begin{pmatrix} A^2 & 0 \\ 0 & I_{p-k} \end{pmatrix}, \qquad \mu = \Sigma^{1/2}u = \begin{pmatrix} v \\ 0 \end{pmatrix}. \tag{20}
$$

With this choice of parameters, $\Delta = \mu^T\Sigma^{-1}\mu = u^Tu = 1$ and $\Delta_1 = \mu_1^T(\Sigma_{11})^{-1}\mu_1 = v^T(\Sigma_{11})^{-1}v = u_1^TA(\Sigma_{11})^{-1}Au_1 = u_1^Tu_1 = 1$, so by (19)

$$
v^TZ_1 = \mu^T\Sigma^{-1/2}Z \overset{d}{=} \mu_1^T(\Sigma_{11})^{-1/2}Z_1 = u_1^TZ_1. \tag{21}
$$

Therefore $u_1^TZ_1$ has the same distribution for any direction in the same half-space as v, and by covering $\mathbb{R}^k$ with overlapping half-spaces, it follows that $u_1^TZ_1$ has a fixed distribution for any direction $u_1 \in S^{k-1}$.

2. Given that Z_1 is spherical, suppose that for any direction $u \in S^{p-1}$, $\mu \in \mathbb{R}^p, \Sigma > 0$ can be found such that $\Sigma^{-1/2}\mu = u$ and $|(\Sigma_{11})^{-1/2}\mu_1| = 1$. Then by (19), $u^TZ \overset{d}{=} \mu_1^T(\Sigma_{11})^{-1/2}Z_1$ where the projection $\mu_1^T(\Sigma_{11})^{-1/2}Z_1$ has a fixed distribution independent of u, the distribution of the LHS likewise does not depend on u, i.e., Z is spherical, and the affine transformations $\{\Sigma^{1/2}Z + \mu : \mu, \Sigma > 0\}$ are elliptic.

A choice of parameters satisfying this requirement is given for the case $u_1 \neq 0$ as

$$\Sigma^{1/2} = \begin{pmatrix} I_k & u_1 u_2^T/(1-|u_2|^2) \\ u_2 u_1^T/(1-|u_2|^2) & C \end{pmatrix}, \qquad \mu = \Sigma^{1/2} u. \tag{22}$$

The lower right block C left unspecified above is chosen given the other blocks of $\Sigma^{1/2}$ to ensure $\Sigma^{1/2} > 0$. For u in the lower dimensional subspace $\{u_1 = 0\}$ the distributions of $u^T Z$ are then determined by continuous mapping, and must be the common distribution of $u^T Z$ for $u_1 \neq 0$.

□

If X, Y are independent, elliptically distributed random vectors with the same variance matrix, then $X - Y$ is elliptically distributed (Cambanis et al., 1981), and taking all affine transformations of $X - Y$ gives an elliptical location-scale family. The theorem then implies that testing for a difference between the full and reduced AUCs, where the controls are distributed as X and cases as Y, is the same as testing for a difference between the full and reduced Mahalanobis distances between X and Y. The control distribution X may even belong to a different ellipitcal family than the case distribution Y, say multivariate normal controls and multivariate t cases, as long as the variances are the same. When both control and case distributions are normal, the homoscedasticity requirement can be dropped. The difference of independent Gaussian vectors is Gaussian and therefore elliptically contoured, whether or not the variances are the same.

2.2 Testing for a difference in LDA AUCs

The benefit of the equivalence (Section 1) between testing for a differece in AUCs and a difference in Mahalanobis distances is that the latter is relatively easy. Demler et al. (2011) shows that this equivalence holds for homoscedastic Gaussian data, and a classical result refers the differences in Mahalanobis distance between Gaussians to an F distribution. Specifically, given observations under the homoscedastic Gaussian model (5), let the empirical Mahalanobis difference be

$$\hat{\Delta} - \hat{\Delta}_1 = (\hat{\mu}_Y - \hat{\mu}_X)^T \hat{\Sigma}^{-1} (\hat{\mu}_Y - \hat{\mu}_X) - (\hat{\mu}_{Y1} - \hat{\mu}_{X1})^T (\hat{\Sigma}_{11})^{-1} (\hat{\mu}_{Y1} - \hat{\mu}_{X1}), \tag{23}$$

where $\hat{\mu}_X$ and $\hat{\mu}_Y$ are sample averages of the controls and cases, and $\hat{\Sigma}$ is the pooled variance estimator. Then (Rao, 1973) under the null hypothesis that $\Delta = \Delta_1$,

$$\frac{(n-p-1)(\hat{\Delta} - \hat{\Delta}_1)}{(p-k)(1+\hat{\Delta}_1)} \sim F_{p-k,n-p-1}. \tag{24}$$

Therefore the test that rejects when the above statistic exceeds the upper $1-\alpha$ quantile of the $F_{p-k,n-p-1}$ distribution is a level α test for the difference in Mahalanobis distances, and so, by the the Equivalence Principle, a level α test for the difference in LDA AUCs.

As we have shown the Equivalence Principle holds more broadly for elliptically contoured data, we formulate a more general test. We use an parameter that is equivalent for our purposes,

$$\psi = \mu_2 - \Sigma_{21} \Sigma^{-1} \mu_1, \tag{25}$$

with sample version

$$\hat{\psi} = \hat{\mu}_2 - \hat{\Sigma}_{21} \hat{\Sigma}^{-1} \hat{\mu}_1. \tag{26}$$

The equivalence of $H_0 : \Delta = \Delta_1$ and $H_0 : \psi = 0$ follows from a block matrix decomposition.

Proposition 1. *Given elliptically contoured distributions F_X and F_Y with means μ_X, μ_Y and common variance Σ, and a sample*

$$X_1, \ldots, X_{n_X} \sim F_X, Y_1, \ldots, Y_{n_Y} \sim F_Y. \tag{27}$$

1. Define

$$\phi : (w, \Sigma) \mapsto w_2 - \Sigma_{21} \Sigma_{11}^{-1} w_1 \tag{28}$$

$$\psi : (w, \Sigma) \mapsto \phi(w, \Sigma) w_1^T \Sigma_{11}^{-1} (\mu_{Y1} - \mu_{X1}). \tag{29}$$

Then as $n_X \to \infty, n_Y \to \infty$,

$$\hat{\psi} - \psi = \frac{1}{n_Y}\sum_{i=1}^{n_Y}(\phi(Y_i - \mu_Y, \Sigma) - \psi(Y_i - \mu_Y, \Sigma)) - \frac{1}{n_X}\sum_{i=1}^{n_X}(\phi(X_i - \mu_X, \Sigma) + \psi(X_i - \mu_X, \Sigma)) \\ + o_P(|\hat{\mu}_{X1} - \mu_{X1}| + |\hat{\mu}_{Y1} - \mu_{Y1}| + |\hat{\Sigma}_{11} - \Sigma_{11}|). \tag{30}$$

2. Assuming F_X and F_Y have 4 finite moments and $n_X/n_Y \to \rho \in (0, \infty)$, $\hat{\psi} - \psi$ is asymptotically normal. A consistent estimator $\hat{Var}(\hat{\psi})$ of its asymptotic variance is the sample variance of

$$\left\{\frac{n}{n_Y}(\phi(Y_i - \hat{\mu}_Y, \hat{\Sigma}) - \psi(Y_i - \hat{\mu}_Y, \hat{\Sigma})), \frac{n}{n_X}(\phi(X_j - \hat{\mu}_X, \hat{\Sigma}) + \psi(X_j - \hat{\mu}_X, \hat{\Sigma}))\right\}_{1 \le i \le n_Y, 1 \le j \le n_X}. \tag{31}$$

3. An asymptotic level α test of $H_0 : \Delta = \Delta_1$ based on the sample (27) *rejects when $\sqrt{n}|\hat{\psi}|/\sqrt{\hat{Var}(\hat{\psi})} > z_{1-\alpha/2}$.*

Proof. The linearization (30) is a first-order Taylor expansion of the test statistic (26) in $\hat{\mu}_X, \hat{\mu}_Y$, and $\hat{\Sigma}$. The moment assumption in the second statement ensures the Taylor remainder is negligible. □

The first part gives an IID sum asymptotically equivalent to the test statistic (26). Provided the data are such that the summands have bounded variances, the CLT then implies that the test statistic $\hat{\psi}$ is asymptotically normal. Furthermore, its asymptotic variance can be estimated using the observed variance of the terms in (31). The third part is the resulting hypothesis test. This test is based on the CLT, and can be easily extended by an appropriate CLT variant to accommodate non-identically distributed data, or as in the next subsection, longitudinal data. The difficulty in this extension is in interpretation when translating the result of the test from Mahalanobis distances back to AUCs, as discussed below.

The F-test (24) has the comparative benefit of being valid for any sample size. In exchange for the loss of finite sample validity, we are no longer resitricted to normally distributed data. The tests are asymptotically equivalent in the homoscedastic Gaussian model, with the variances of the respective test statistics each tending to

$$(\rho + 1/\rho)(\Delta_1 + 1)(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}). \tag{32}$$

The finite-sample performance of the proposed test is examined in (sec:sim).

2.3 Longitudinal data

Diagnostic data is often clustered. For example, blood pressure and cholesterol readings may be collected on subjects in a series of visits as a diagnostic for heart disease. In this case, besides the dependence in the measurements from a given visit, we can expect dependence across a given subject's visits. It is also possible that the same individual contributes both control and case readings, such as when a subject changes disease status in the course of a study. To model clustered data, suppose the control and case observations above (27) are listed in a fixed way, and $g_i, 1 \le i \le n = n_x + n_y$, indicates which of G_n clusters each observation belongs to, $1 \le g_i \le G_n$:

$$\begin{aligned} X_i \sim F_X, Y_j \sim F_Y, 1 \le i \le n_X, 1 \le j \le n_Y, &\text{ with } F_X, F_Y \text{ elliptically contoured} \\ E(X_i) = \mu_X, E(Y_i) = \mu_Y, Var(X_i) &= Var(Y_i) = \Sigma \\ \begin{matrix} X_1 & \dots & X_{n_X} & Y_1 & \dots & Y_{n_Y} \\ g_1 & \cdots & g_{n_X} & g_{n_X+1} & \cdots & g_{n_X+n_Y}. \end{matrix} \end{aligned} \tag{33}$$

Cluster $i, 1 \le i \le G_n$, has $n_i = \sum_j \{g_j = i\}$ observations. Observations belonging to the same cluster may depend on each other in arbitrary ways, while the clusters are independent of each other. We assume that the sizes of the clusters are uniformly bounded so that he number of clusters G_n grows with n.

Prop. 2 extends the test in Prop. 1 to accommodate clustered data of the type introduced.

Proposition 2. *1. Suppose data given as in* (33) *has 4 moments and there is a number C such that $n_g < C, 1 \leq g \leq G$. Let*

$$W_i = \begin{cases} -\frac{n}{n_X}(\phi(X_j - \mu_X, \Sigma) + \psi(X_j - \mu_X, \Sigma)) & 1 \leq i \leq n_X \\ \frac{n}{n_Y}(\phi(Y_i - \mu_Y, \Sigma) - \psi(Y_i - \mu_Y, \Sigma)) & n_X + 1 \leq i \leq n \end{cases} \tag{34}$$

Then $\hat{\psi} - \psi$ is asymptotically normal with asymptotic variance given by $\frac{1}{n}\sum_{i=1}^{G_n} E\left(\sum_{j:g_j=i} W_j W_J^T\right)$.

2. Let $\hat{Var}(\hat{\psi}) = \frac{1}{n}\sum_{i=1}^{G_n} E\left(\sum_{j:g_j=i} \hat{W}_j \hat{W}_J^T\right)$, where $\hat{W}_i, i = 1, \ldots, n$, are formed as in (34) *but using $\sqrt{n}-$consistent estimators in place of μ_X, μ_Y, Σ. An asymptotic level α test of $H_0 : \Delta = \Delta_1$ based on the sample* (33) *rejects when $\sqrt{n}|\hat{\psi}|/\sqrt{\hat{Var}(\hat{\psi})} > z_{1-\alpha/2}$.*

Proof. The proposition follows from the linearization (30) and a cluster-robust CLT, such as given in Hansen and Lee (2019). □

As before, if data is elliptically contoured, the test given in Prop. 2 is a valid test for a difference in LDA AUCs. If the data is also Gaussian, LDA is the optimal decision rule, and the test is for a difference in the best diagnostics based on the data.

We comment on the assumptions. The cluster sizes are assumed to be bounded. A typical example is the longitudinal study where many patients have their measurements observed over a fixed period of time. This assumption can easily be relaxed to allow for unbounded cluster sizes, as long as the cluster sizes don't grow too fast relative to the total number of observations. The cost would be higher moment requirements than the 4 assumed. An altogether different approach would be required for data where the cluster sizes grow at the same or greater order than the total number of observations, such as if a fixed number of patients were observed for an open-ended stretch of time.

Second, the reqirement that all observations have the same first 2 moments may appear restrictive. In fact, this condition may be relaxed by using a triangular array CLT. The difficulty is in the interpretation of the test of the null $H_0 : \Delta = \Delta_1$. In the presence of dependence there are several concepts that can lay a claim to being the proper generalization of the AUC (Michael et al., 2019; Michael and Tian, 2024). These coincide when the first two moments are fixed.

3 Simulation

We examine the performance of the tests proposed in Section 2. We observe the rejection rate at the null, assessing FPR control, and known alternatives, assessing power. We test both normal and nonnormal, elliptically contoured data using the test in Prop. 1, and clustered and unclustered data using the test in Prop. 2.

3.1 Data generation

A data set is generated by first sampling n_X control and n_Y case observations of dimension p under F_X and F_Y, where F_X and F_Y are elliptically contoured with means μ_X and μ_Y, and common variance Σ. As in (33), the observations are arranged in a fixed order, and cluster IDs $g_i, 1 \leq g_i \leq G, 1 \leq i \leq n$, indicating which of G clusters an observation belongs to, are assigned randomly to the vector of observations. Next, G p-dimensional vectors $Z_1, \ldots, Z_G$, are sampled with mean 0 and variance $\sigma_\lambda I$. Each of these effects is added to all elements in the corresponding cluster, inducing dependence within the clusters, i.e., for all i such that $g_i = j$, Z_j is added to observation i. The marginal variance of each observation is therefore $\Sigma + \sigma_\lambda I$. The marginal means and variance are related through

$$\mu_{Y2} - \mu_{X2} = \psi + \Sigma_{21}\Sigma_{11}^{-1}(\mu_{Y1} - \mu_{X1}) \tag{35}$$

for $\psi \in \mathbb{R}^{p-k}$. As defined in (25), $\psi = 0$ corresponds to the null case $\Delta = \Delta_1$, and $\psi \neq 0$ to alternatives.

3.2 Parameters

We take $p = 5$ and $k = 4$, so that the diagnostic effect of $p - k = 1$ additional covariate is under consideration. The control and case distributions F_X, F_Y are taken to be multivariate Gaussians and multivariate t's.

We set $n_X = 50, 200, 1000$, $n_Y = 50, 200, 1000$, $n = n_X + n_Y$, partitioned among $G = 2, 4, 40$ groups of equal size. Values of ψ ranging between 0 and .1 are chosen. The parameters Σ and μ_{Y1} are chosen randomly, Σ is scaled to unit variances, μ_X is set to 0, and μ_Y is set for each value of ψ by (35). The random effect variance is chosen to induce within-cluster correlation of magnitude $\sigma_\lambda^2/(1+\sigma_\lambda^2) = 0, .3, .6$.

To interpret the alternatives given as Mahalanobis differences they are converted to the difference in LDA AUC, which is the parameter of interest. For Gaussian F_X and F_Y, the formula given in (7) can be used, which maps Mahalanobis distance to AUC as $\Delta \mapsto \Phi(\Delta/\sqrt{2})$. When F_X and F_Y are multivariate t with $\nu > 2$ degrees of freedom, the mapping from Mahalanobis distance to AUC is

$$\Delta \mapsto \frac{1}{2} - \frac{1}{\pi}\frac{((\nu-2)\Delta)^{\nu/2}}{2^{\nu-2}\Gamma(\nu/2)^2}\int_0^\infty t^{\nu-1}K_{\nu/2}(t\sqrt{(\nu-2)\Delta})^2\sin(t\Delta)dt,$$

where K is the modified Bessel function of the second kind.

Proof. The p-dimensional multivariate t distribution can be parametrized by $\nu > 0$, representing degrees of freedom, mean $\mu \in \mathbb{R}^p$, and a positive definite $p \times p$ matrix Λ related to the variance through $\Sigma = \frac{\nu}{\nu-2}\Lambda$. Suppose X and Y are multivariate t with means μ_X and μ_Y, common degrees of freedom ν, and common scale matrix Λ, and let $\beta_{LDA} = \Sigma^{-1}(\mu_Y - \mu_X)$ as before. The characteristic function of X is (Joarder and Ali, 1996),

$$\phi_X : t \mapsto e^{it^T\mu_X}\frac{|(\nu\Lambda)^{1/2}t|^{\nu/2}}{2^{\nu/2-1}\Gamma(\nu/2)}K_{\nu/2}(|(\nu\Lambda)^{1/2}t|), \tag{36}$$

and analogously for Y. By a Fourier inversion formula (Gil-Pelaez, 1951), the AUC between X and Y is then

$$P(\beta_{LDA}{}^TX < \beta_{LDA}{}^TY) = \frac{1}{2} - \frac{1}{\pi}\int_0^\infty t^{-1}\Im(\phi_{\beta_{LDA}{}^T(X-Y)}(t))dt \tag{37}$$

$$\frac{1}{2} - \frac{1}{\pi}\int_0^\infty t^{-1}\Im(\phi_X(t\beta_{LDA})\phi_Y(-t\beta_{LDA}))dt. \tag{38}$$

Since $|(\nu\Lambda)^{1/2}t\beta_{LDA}| = |t|\sqrt{(\nu-2)\Delta}$ and $e^{it\beta_{LDA}{}^T(\mu_X-\mu_Y)} = e^{-it\Delta}$, (38) is

$$\frac{1}{2} - \frac{1}{\pi}\int_0^\infty t^{-1}\frac{(t\sqrt{(\nu-2)\Delta})^\nu}{2^{\nu-2}\Gamma(\nu/2)^2}K_{\nu/2}(t\sqrt{(\nu-2)\Delta})^2\Im(e^{it\beta_{LDA}{}^T\mu_X}e^{-it\beta_{LDA}{}^T\mu_Y})dt \tag{39}$$

$$\frac{1}{2} - \frac{1}{\pi}\frac{((\nu-2)\Delta)^{\nu/2}}{2^{\nu-2}\Gamma(\nu/2)^2}\int_0^\infty t^{\nu-1}K_{\nu/2}(t\sqrt{(\nu-2)\Delta})^2\sin(t\Delta)dt. \tag{40}$$

□

In this way 500 data sets were constructed for each combination of parameter settings. The test of the null $H_0 : \Delta = \Delta_1$ given in Prop. 2 was then applied to each, as well as the tests from Demler et al. (2011) and DeLong et al. (1988). Whether a given test rejects or not was recorded and averaged over the data sets to estimate the FPR ($\psi = 0$) and power ($\psi \neq 0$) of the tests.

3.3 Results

Results are given in Figs. 1 and 2 for Gaussian and t data. In the matrix of plots, sample size increases along the rows and within-cluster correlation along the columns. The last two columns indicate that the exact test does not control the FPR, which is unsurprising as the test's independence assumption is violated. The Delong test is underpowered throughout, which is also expected from previous analyses (see Section 1) when the Delong test is applied to fitted values. The bottom row, where sample size is large, indicates that the proposed test controls the FPR while maintaining a power advantage over Delong test. The proposed test is based on asymptotics however, and for smaller sample sizes suffers from the presence of dependence much as the exact test does.

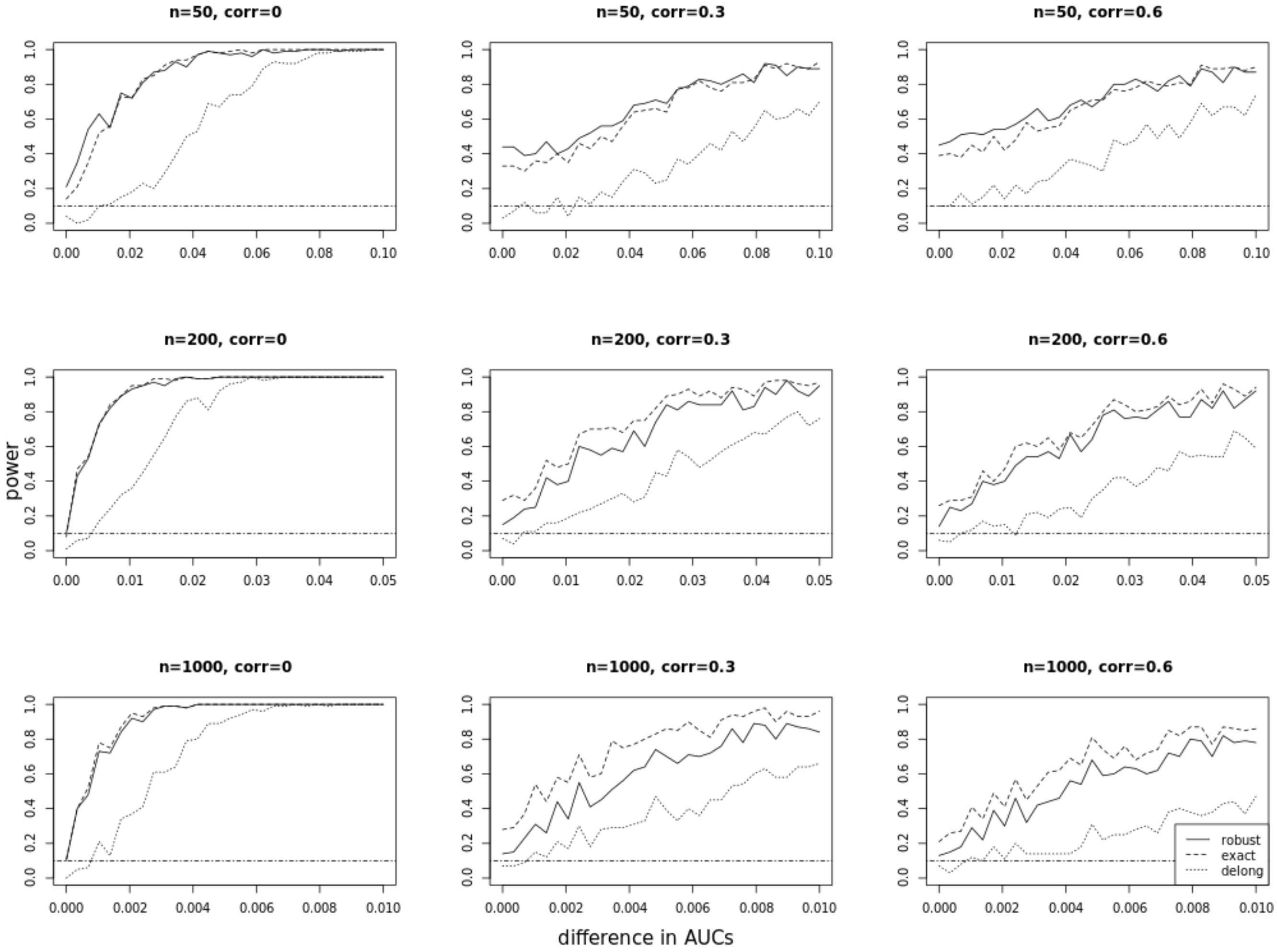

Figure 1: Power curves for multivariate Gaussian data.

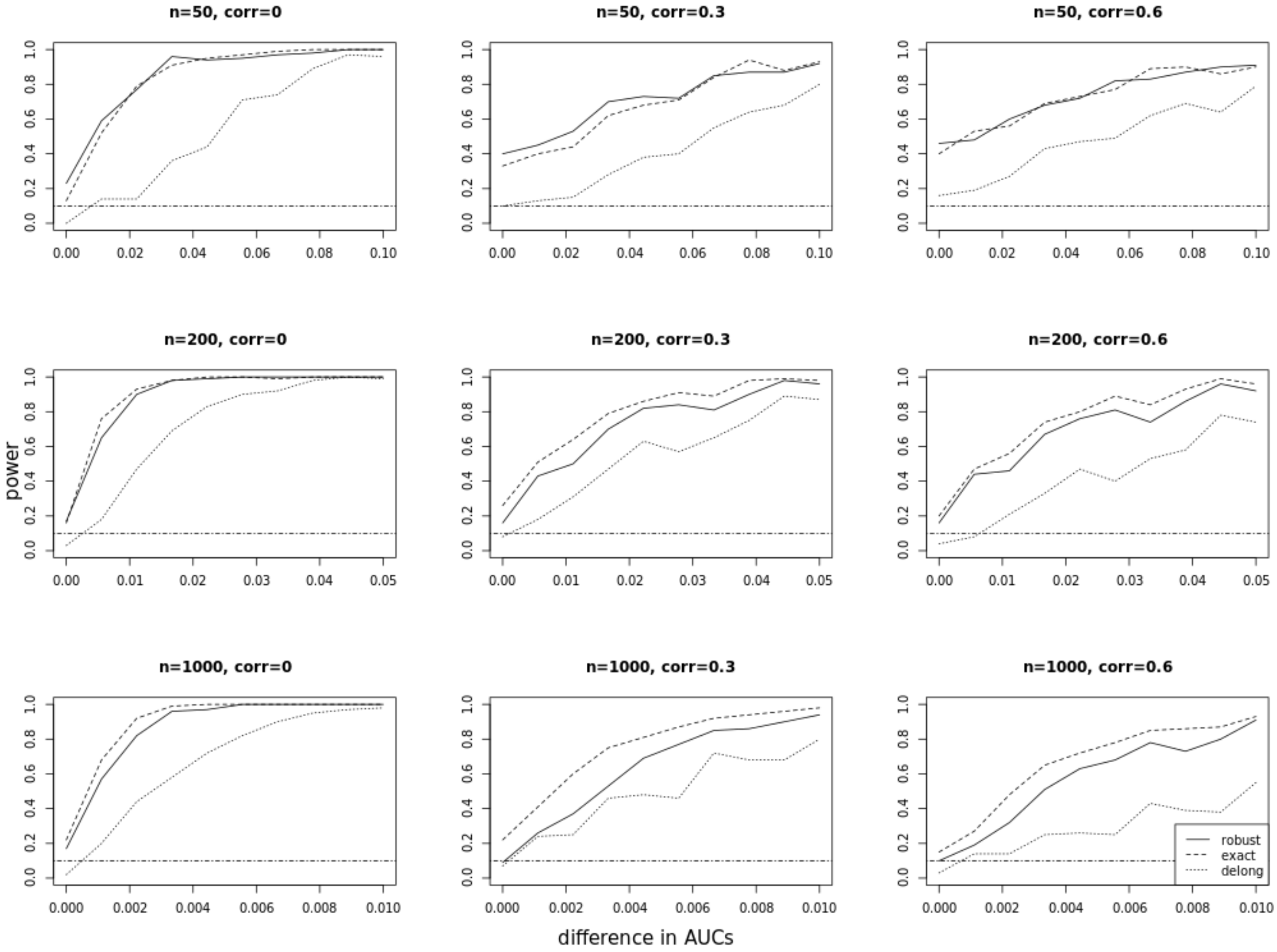

Figure 2: Power curves for multivariate t data.

4 Conclusion

We have developed a test for the difference in AUCs using elliptically contoured data. An important avenue for further work is to improve the efficiency of the currently proposed estimator. It is known that the family of elliptically contoured distributions admits adaptive estimators for any parameter that is a function of the mean and variance and homogeneous of degree 0 in the variance, a class that includes the Mahalanobis parameter used above. That is, it is theoretically possible to construct an estimator for any elliptically contoured data that achieves the same asymptotic variance as an efficient estimator taking into account the specific parametric form of the data. Such an estimator may improve the performance of the proposed esetimator in smaller samples or with high dependency.

References

Cambanis, S., S. Huang, and G. Simons (1981). On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis 11*(3), 368–385.

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-Connolley, A. and H. Michael (2024). Nonparametric estimation of the AUC of an index with estimated parameters. Forthcoming; available at https://www.umass.edu/mathematics-statistics/directory/faculty/haben-michael.

Gil-Pelaez, J. (1951). Note on the inversion theorem. *Biometrika 38*(3-4), 481–482.

Hansen, B. E. and S. Lee (2019). Asymptotic theory for clustered samples. *Journal of econometrics 210*(2), 268–290.

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.

Joarder, A. H. and M. M. Ali (1996). On the characteristic function of the multivariate t-distribution. *Pakistan Journal of Statistics 12*, 55–62.

Michael, H. et al. (2024). Inference on the difference of index AUCs under the null. Forthcoming; available at https://www.umass.edu/mathematics-statistics/directory/faculty/haben-michael.

Michael, H. and L. Tian (2024). The population and personalized AUCs. Forthcoming; available at https://www.umass.edu/mathematics-statistics/directory/faculty/haben-michael.

Michael, H., L. Tian, and M. Ghebremichael (2019). The roc curve for regularly measured longitudinal biomarkers. *Biostatistics 20*(3), 433–451.

Rao, C. R. (1973). *Linear statistical inference and its applications*, Volume 2.

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

Su, J. Q. and J. S. Liu (1993). Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association 88*(424), 1350–1355.

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.