The data consists of clustered control and case pairs of real-valued markers:

$$(X_{ij}, X'_{ij}), i = 1, \ldots, m_i, (Y_{ij}, Y'_{ij}), i = 1, \ldots, n_i, i = 1, \ldots, N.$$

In our application, $(X_{ij}, X'_{ij}), j = 1, \ldots, m_i$, represent, for given subject i, the subject's m_i non-diseased CD4 and CD4P markers, and $(Y_{ij}, Y'_{ij}), j = 1, \ldots, n_i$, represent his diseased CD4 and CD4P markers. N represents the number of subjects. The vectors $(\{(X_{ij}, X'_{ij})\}_{i=1}^{m_i}, \{(Y_{ij}, Y'_{ij})\}_{i=1}^{m_i}, m_i, n_i), j = 1, \ldots, N$, are assumed to be independent, i.e., there is no dependency between clusters. On the other hand, there may be dependency within a cluster, e.g., among diseased or non-diseased observations or between the markers. [[depenency may be longitudinal or other time]]

We seek to adapt the AUC to this setting, quantifying the extent to which diseased observations (Y_{ij}, Y'_{ij}) tend to exceed non-diseased observations (X_{ij}, X'_{ij}). Doing so requires some refinement to the definition of the AUC. One statistic for measuring the AUC given by the control and case measurements X_{ij}, Y_{ij} averages the Mann-Whitney non-parametric AUC estimate of the clusters:

$$U_N = \frac{1}{N} \sum_{i=1}^{N} H_{ii} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i m_i} \sum_{j,k} \{X_{ij} < Y_{ik}\}, . \tag{1}$$

Here $H_{ij} := \frac{1}{n_i m_j} [[fixm/n]] \sum_{j,k} \{X_{ij} < Y_{ik}\}$ is the AUC statistic computed using the non-diseased measurements of cluster i and the diseased measurements of cluster j. Another statistic computes the non-parametric estimate ignoring the cluster structure,

$$V_N = \frac{1}{N^2} \sum_{i,j} H_{ij} = \frac{1}{\sum m_i \sum n_i} \sum_{i,j} \sum_{r,s} \{X_{ir} < Y_{js}\}. \tag{2}$$

U_N represents the AUC of a typical patient in the population, whereas V_N represents the population AUC, marginalizing over patients. As an illustration of the difference, suppose $\theta_1, \ldots, \theta_N$ are uniformly distributed on $(0, 1) \subset \mathbb{R}$. Conditionally on θ_i, let patient i's non-diseased observations be uniformly distributed on $(\theta_i - 1, \theta_i)$ and his diseased observations uniformly distributed on $(\theta_i, \theta_i + 1)$. Then each patient's non-diseased and diseased observations are perfectly separated, so the average individual AUC is 1. On the other hand, marginally as θ_i varies, there is no such perfect separation, e.g., non-diseased observations in the population are $> .5$ with positive probability and diseased observations are $< .5$ with positive probability. So the population AUC is < 1. The population AUC is probably more commonly studied, e.g., [3], [2].

U_N and V_N are similar to U-statistics, but the dependency in the data prevents a straightforward application of U-statistics theory. We try to avoid making strict parametric assumptions on the complicated dependency, instead prefering general assumptions such as weak dependency.

1 Individual AUC

First suppose the clusters, besides being mutually independent, are also identically distributed:

$$(\{(X_{ij}, X'_{ij})\}_{i=1}^{m_i}, \{(Y_{ij}, Y'_{ij})\}_{i=1}^{m_i}, m_i, n_i), j = 1, \ldots, N, \text{ are iid.}$$

Then (U_N, U'_N) is a sum of iid variables of bounded variance. The prime indicates a quantity computed on the second set of biomarkers (X'_{ij}, Y'_{ij}). Let $\theta_{11} = \mathbb{E}[U_N] = \mathbb{E}[H_{11}] = \mathbb{P}[X_{11} < Y_{11}]$ denote the probability that a non-diseased value is less than a diseased value in the same cluster,

which is well-defined by the iid assumption, and analogously $\theta'_{11} = \mathbb{E}[U'_N] = \mathbb{E}[H'_{11}] = \mathbb{P}[X'_{11} < Y'_{11}]$. By the CLT,

$$Cov(H_{11}, H'_{11})^{-1/2}(\sum_{i=1}^{N}(H_{ii}, H'_{ii}) - N(\theta_{11}, \theta'_{11})) \rightsquigarrow \mathcal{N}(0, I),$$

and so,

$$\sqrt{N}(U_N - \theta_{11}) \rightsquigarrow \mathcal{N}(0, Cov(H_{11}, H'_{11})).$$

This covariance, whatever its form, may be estimated consistently from the data. We can obtain its form in specific cases. For example, suppose the observations within a cluster are exchangeable [[also describe corr bw markers]]. More generally, suppose that within any cluster i the following quantities do not vary with j, k, l, m:

$$P(X_{ij} < Y_{ik}), P(X_{ij} < Y_{ik}, X_{ij} < Y_{il}), P(X_{ik} < Y_{ij}, X_{il} < Y_{ij}), P(X_{ij} < Y_{ik}, X_{il} < Y_{im}). \quad (3)$$

This assumption applies, for example, to the situation in which a generalized linear mixed effects model relates disease status S_{ij} to some marker U_{ij},

$$P(S_{ij} = 1) = g(\alpha_i + \beta_i U_{ij}), j = 1, \ldots, r_i, i = 1, \ldots, N,$$

the U_{ij} are divided by binary status to give the first set of markers (e.g., CD4),

$$\{X_{ij}\}_{i=1}^{m_i} = \{U_{ij} | S_{ij} = 0, j = 1, \ldots, r_i\}, \{Y_{ij}\}_{i=1}^{n_i} = \{U_{ij} | S_{ij} = 1, j = 1, \ldots, r_i\}, m_i + n_i = r_i,$$

and the second set of markers (e.g., CD4P) are obtained as noisy surrogates of the first, with the noise level depending on subject,

$$\begin{aligned} U'_{ij} &= U_{ij} + \gamma_i \epsilon_{ij}, \epsilon_{ij} \sim \mathcal{N}(0, 1) \\ \{X'_{ij}\}_{i=1}^{m_i} &= \{U'_{ij} | S_{ij} = 0, j = 1, \ldots, r_i\}, \{Y'_{ij}\}_{i=1}^{n_i} = \{U'_{ij} | S_{ij} = 1, j = 1, \ldots, r_i\}. \end{aligned}$$

Let the vector lengths m_i and n_i be independent of the vector values $(\{(X_{ij}, X'_{ij})\}_{i=1}^{m_i}, \{(Y_{ij}, Y'_{ij})\}_{i=1}^{m_i})$, and let $m = E(m_1), n = E(n_1)$. Under the assumption (3) the asymptotic variance $Cov(H_{11}, H'_{11})$ is computed as

$$\begin{aligned} Cov(H_{11}, H'_{11})_{11} =& \frac{1}{mn}(\mathbb{P}[X_{i1} < Y_{i1}] + (n-1)\mathbb{P}[X_{i1} < Y_{i1}, X_{i1} < Y_{i2}] + (m-1)\mathbb{P}[X_{i1} < Y_{i1}, X_{i2} < Y_{i1}] \\ &+ (n-1)(m-1)\mathbb{P}[X_{i1} < Y_{i1}, X_{i2} < Y_{i2}]) - \mathbb{P}^2[X_{i1} < Y_{i1}] \\ Cov(H_{11}, H'_{11})_{12} =& \frac{1}{mn}(\mathbb{P}[X_{i1} < Y_{i1}, X'_{i1} < Y'_{i1}] + (n-1)\mathbb{P}[X_{i1} < Y_{i1}, X'_{i1} < Y'_{i2}] \\ &+ (m-1)\mathbb{P}[X_{i1} < Y_{i1}, X'_{i2} < Y'_{i1}] + (n-1)(m-1)\mathbb{P}[X_{i1} < Y_{i1}, X'_{i2} < Y'_{i2}]) \\ &- \mathbb{P}[X_{i1} < Y_{i1}]\mathbb{P}[X'_{i1} < Y'_{i1}] \\ Cov(H_{11}, H'_{11})_{22} =& \frac{1}{mn}(\mathbb{P}[X'_{i1} < Y'_{i1}] + (n-1)\mathbb{P}[X'_{i1} < Y'_{i1}, X'_{i1} < Y'_{i2}] + (m-1)\mathbb{P}[X'_{i1} < Y'_{i1}, X'_{i2} < Y'_{i1}] \\ &+ (n-1)(m-1)\mathbb{P}[X'_{i1} < Y'_{i1}, X'_{i2} < Y'_{i2}]) - \mathbb{P}^2[X'_{i1} < Y'_{i1}]. \end{aligned}$$

Figure 1 displays the results of simulations on synthetic data with $N = 30$ patients.

Again, since the clusters are assumed iid, even without an assumption such as (3) we may estimate the asymptotic variance by usual sample variance estimators. The explicit formula given above may be useful in cases where the concordance parameters (3) are already known.

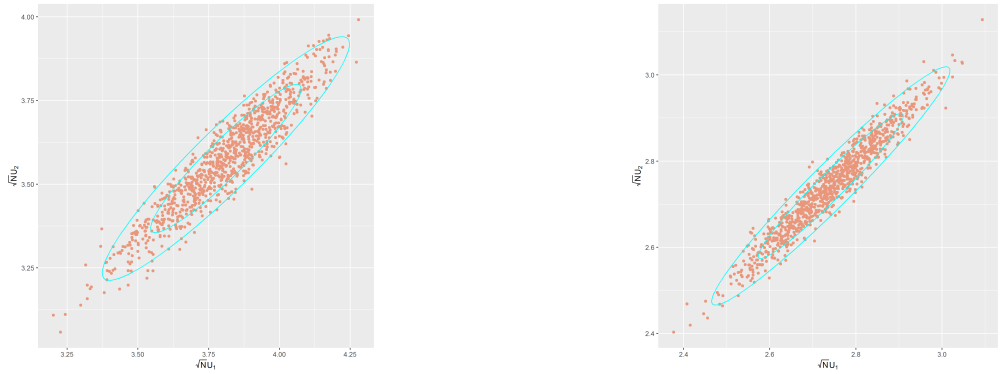

Figure 1: Scatterplot of 1e3 realizations of (U_N, U'_N), along with 67% and 95% level curves of the $N((\hat{\theta}_{11}, \hat{\theta}_{12}), \hat{\Sigma})$ density. For the left figure, the data in each replicate consists of $N = 30$ pairs $(X_i, Y_i), (X'_i, Y'_i)$ of correlated vectors generated by a GLMM model, with (X'_i, Y'_i), the marker measured by U'_N, being a noisy reading of (X_i, Y_i), the marker measured by U_N. For the right figure the biomarkers are marginally uniform on $[0, 1]$ with a gaussian copula accounting for covariance.

1.1 Application to HIV data

We examine the AUC of a typical subject in the Yale HIV pediatric data, using both CD4 and CD4P as predictors of blip status. We find that the individual AUCs are approximately normal with parameters:

$$(AUC_{CD4}, AUC_{CD4P})^T \dot{\sim} \mathcal{N}((.6597, .6623)^T, \begin{pmatrix} .0030 & .0021 \\ .0021 & .0036 \end{pmatrix}),$$

giving a contrast:

$$AUC_{CD4P} - AUC_{CD4} \dot{\sim} \mathcal{N}(-.0026, .0005)$$

and 95% CI of (-.0210,.0157), giving no reason to reject the null of no difference between the individual AUCs.

2 Population AUC

We next consider the population (marginal) AUC (2),

$$(V_N, V'_N) = \frac{1}{N^2}(\sum_{i,j} H_{ij}, \sum_{i,j} H'_{ij}), \tag{4}$$

where as before the primed quantities refer to the second set of biomarkers (e.g., CD4P). Assuming that the clusters are independent (not necessarily identically distributed), the population AUC (V_N, V'_N) is asymptotically normal:

$$\sqrt{N}\Sigma^{-1/2} \begin{pmatrix} Var(V_N)^{-1/2}(V_N - \mathbb{E}(V_N)) \\ Var(V'_N)^{-1/2}(V'_N - \mathbb{E}(V'_N)) \end{pmatrix} \rightsquigarrow \mathcal{N}(0, I). \tag{5}$$

The mean is asymptotically equal to

$$\mathbb{E}[V_N] \dot{\sim} \sum_{i \neq j} \mathbb{E}[H_{ij}],$$

and the covariance matrix $Var(V_N)$ to

$$\begin{aligned} Var(V_N)_{11} &\dot{\sim} (N(N-1))^{-1} \sum_{|\{i,j,k,l\}|=3} Cov[H_{ij}, H_{kl}] \\ Var(V_N)_{12} = Var(V_N)_{21} &\dot{\sim} (N(N-1))^{-1} \sum_{|\{i,j,k,l\}|=3} Cov[H_{ij}, H'_{kl}] \\ Var(V_N)_{22} &\dot{\sim} (N(N-1))^{-1} \sum_{|\{i,j,k,l\}|=3} Cov[H'_{ij}, H'_{kl}]. \end{aligned}$$

The notation $|\{i,j,k,l\}| = 3$ indicates the summation is over all sets of 4 indexes $1 \leq i,j,k,l \leq N$ of which exactly 3 are distinct.

Asymptotic normality holds without any further assumptions besides independence of the clusters. Therefore this result generalizes [2], which assumes exchangeability. [[also a generalizatoin of obuchowski, roser, sen]]. The result also extends [2] in two other ways. First, [2] assumes independence between all case and control observations, in addition to independence of the clusters. We are not free to make this assumption because in our application a given subject contributes both case and control observations. Second, we consider 2 biomarkers, the primed observtions (X'_{ij}, Y'_{ij}) in addition to the unprimed observations (X_{ij}, Y_{ij}), so that we can compare the AUCs of the biomarkers. The result extends the foundational work on U-statistics of dependent data [4] by [[dropping stationarity assumption, vector valued, etc]].

The demonstration starts by writing V_N as

$$V_N = \frac{1}{N^2} \sum_{i=1}^{N} H_{ii} + \frac{N(N-1)}{N^2} W_N,$$

defining the statistic $W_N = 1/(N(N-1)) \sum_{i \neq j} H_{ij}$ as the average of the AUCs where the non-diseased data comes from one cluster and the diseased from another cluster. The first term on the rhs is asymptotically negligible, so the asymptotic distribution of V_N is the same as that of W_N:

$$\begin{aligned} \sqrt{N}(V_N - \mathbb{E}[V_N]) &= N^{-3/2} \sum_{i=1}^{N} H_{ii} + \sqrt{N} \frac{N(N-1)}{N^2} (W_N - \frac{1}{N(N-1)} \sum_{i \neq j} \mathbb{E}[H_{ij}]) + N^{-3/2} \sum_{i=1}^{N} \mathbb{E}[H_{ii}] \\ &\dot{\sim} \sqrt{N}(W_N - \mathbb{E}[W_N]). \end{aligned}$$

To obtain the asymptotic distribution of W_N, we use a commonly used technique to remove the dependency among the summands of a statistic such as W_N, the "Hajék projection," obtaining a projection that is an iid sum and then establishing that the projection and W_N have the same limiting distribution. In more detail, W_N is projected onto the σ-algebra generated by sums of borel-measurable functions of the independent clusters,

$$\begin{aligned} \hat{W}_N &:= \mathbb{E}[W_N \mid \sigma(\{\sum_{i=1}^{N} g_i(X_i) + h_i(Y_i) : g_i, h_i \in \mathcal{B}(\mathbb{R})\})] \\ &= \sum_{i=1}^{N} \mathbb{E}[W|(X_i, Y_i)] - (N-1)\mathbb{E}[W_N]. \end{aligned}$$

See, e.g., [5], chapter 11, for details. One proceeds by first showing that $\mathbb{E}[(W_N - \hat{W}_N)^2] = O(1/N^2)$. Then $\sqrt{N}(W_N - \hat{W}_N) = o_P(1)$ and so the asymptotic distribution of $\sqrt{N}(W_N - \mathbb{E}[W_N])$ may be obtained using that of $\sqrt{N}(\hat{W}_N - \mathbb{E}[\hat{W}_N])$. Since $\hat{W}_N$ is a sum of independent random variables, its asymptotic distribution can be computed using a triangular array CLT. The Lindeberg condition can be applied directly, or, using boundedness of the terms H_{ij}, which are proportions, simpler corollaries of the Lindeberg-Feller CLT (e.g., [1], Ex. 27.4). The same procedure applies to obtain the asymptotic normality of V'_N, and the joint asymptotic normality of (V_N, V'_N) may be obtained using the Cramér-Wold device.

TODO: The asymptotic normality in (5) is only useful when the asymptotic mean and variance converge. Without any assumptions on the data, as above, sequences of patient data can be constructed such that the asymptotic mean and/or variance diverge. Therefore, appropriate conditions on the data need to be found, rather than the completely nonparametric problem described above.

For example, we may recover the results of [2] by applying (5) under the authors' assumptions 1) that all clusters are independent, 2) that diseased and non-diseased observations belonging to the same patient are independent, 3) the non-diseased observations belonging to a patient are exchangeable, and similarly for the diseased observations, 4) the marginal distribution of non-diseased observations is the same for all patients and all times, and similarly for diseased observations, and 5) the number of non-diseased observations of a given patient, $m_i, i = 1, \ldots, m$, is bounded above, where we ignore patients with no non-diseased observations in this indexing and m is the number of patients with at least one non-diseased observation, and likewise the number of diseased observations of a given patient, $n_i, i = 1, \ldots, n$, is bounded above, where n is the number of patients with at least one diseased observation. The authors focus not directly on the population AUC $V_N = N^{-2} \sum_{i,j} H_{ij}$ but on the Mann-Whitney rank sum statistic,

$$\sum_{i=1}^{m} \sum_{j=1}^{n} m_i n_j H_{ij} = \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{m_i} \sum_{l=1}^{n_j} \{X_{ik} < Y_{jl}\},$$

which differs by the normalization.

By assumptions 1 and 4, $Cov[\{X_{ir} < Y_{js}\}\{X_{iu} < Y_{kv}\}] = Cov[\bar{F}_X(Y_{11}), \bar{F}_X(Y_{12})]$ for all $j \neq k, r, s, u, v$, where F_X is the marginal distribution of the non-diseased observations under assumption 4 and $\bar{F}_X = 1 - F_X$. Denote this quantity r_x following the notation of [2]. Similarly, $Cov[\{X_{jr} < Y_{is}\}\{X_{ku} < Y_{iv}\}] = Cov[F_Y(Y_{11}), F_Y(Y_{12})] := r_y, Var[F_X(Y_{rs})] =: v_y, Var[F_Y(X_{rs})] =: v_x$, for all $j \neq k, r, s, u, v$. The asymptotic variance given in (5) simplifies to:

$$\begin{aligned}
&(mn(m-1)(n-1))^{-1} \sum_{|\{i,j,k,l\}|=3} Cov[H_{ij}, H_{kl}] \\
&= (mn(m-1)(n-1))^{-1} \Big(\sum_{i=1}^{m} \sum_{1 \leq j \neq k \leq n} m_i^2 n_j n_k \Big(\frac{m_i - 1}{m_i} r_x + \frac{1}{m_i} v_x\Big) + \sum_{1 \leq j \neq k \leq m} \sum_{i=1}^{n} m_j m_k n_i^2 \Big(\frac{n_i - 1}{n_i} r_y + \frac{1}{n_i} v_y\Big)\Big) \\
&= (mn(m-1)(n-1))^{-1} \Big(\sum_{1 \leq j \leq k \leq n} n_j n_k (m\bar{m}^{(2)} r_x + m\bar{n} v_x) + \sum_{1 \leq j \leq k \leq m} m_j m_k (n\bar{n}^{(2)} r_y + n\bar{n} v_y)\Big) \\
&\dot{\sim} \frac{1}{m} \bar{n}^2 (\bar{m}^{(2)} r_x + \bar{m} v_x) + \frac{1}{n} \bar{m}^2 (\bar{n}^{(2)} r_y + \bar{n} v_y),
\end{aligned}$$

as given by Theorem 3 of [2]. We follow the authors' notation $\bar{m} = (\sum_{i=1}^{m} m_i)/m, \bar{m}^{(2)} = \sum_{i=1}^{m} (m_i^2 - m_i)$, and analogously for $\bar{n}, \bar{n}^{(2)}$.

[[make var/cov typography even, brackets vs parens]]

References

[1] Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 2008.

[2] Mei-Ling Ting Lee and Herold G. Dehling. Generalized two-sample U-statistics for clustered data. *Statistica Neerlandica*, 59(3):313–323, 2005.

[3] Nancy Obuchowski. Nonparametric analysis of clustered ROC curve data. *Biometrics*, 53:567–578, 1997.

[4] Pranab Kumar Sen. On the properties of U-statistics when the observations are not independent. *Calcutta Statistical Association Bulletin*, 12(3):69–92, 1963.

[5] A. W. van der Vaart. *Asymptotic Statistics:*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.