APPENDIX

A.1. Connection between U-Index and Area Under the ROC Curve

Let $F(g)$, $F_D(g)$ and $F_{\bar{D}}(g)$ denote the c.d.f of ordered genotype g, $g \in \{g_1, \ldots\ldots, g_K\}$ in the entire populations, diseased population and non-diseased population, respectively, so that

$$F(g) = P(G \le g)\text{; } F_D(g) = P(G \le g \mid D)\text{; and } F_{\bar{D}}(g) = P(G \le g \mid \bar{D})$$

A classification rule can be formed by using a particular multi-locus genotype g as threshold:

$$\hat{Y}_G = \begin{cases} 1 & r_G > r_g \\ 0 & r_G \le r_g \end{cases}$$

The ROC curve can then be represented by a map: $f: t \to f(t)$, so that

$$t_g = 1 - P(\hat{Y}_g = 0 \mid \bar{D}) = 1 - F_{\bar{D}}(g)\text{; and } f(t_g) = P(Y_g = 1 \mid D) = 1 - F_D(g)\,.$$

On the other hand, the predictiveness curve can be represented by a map: $r: q \to r(q)$, so that

$$q = F(g)\text{; and } r(q) = P(D \mid g)$$

Let $F'(g)$ and $F'_{\bar{D}}(g)$ be the p.d.f of ordered genotype g in the diseased and non-disease populations, respectively, we would have

$$f'(t_g) = \frac{df(t_g)}{dt_g} = \frac{F'_D(g)}{F'_{\bar{D}}(g)}\,.$$

It follows then:

$$q = F(g) = \rho F_D(g) + (1-\rho) F_{\bar{D}}(g) = \rho(1 - f(t_g)) + (1-\rho)(1-t_g)$$

$$\begin{aligned} r(q) &= P(D \mid g) \\ &= \frac{P(g \mid D)P(D)}{P(g \mid D)P(D) + P(g \mid \bar{D})P(\bar{D})} \\ &= \frac{F'_D(g)\rho}{F'_D(g)\rho + F'_{\bar{D}}(g)(1-\rho)} \\ &= \frac{f'(t_g)\rho}{f'(t_g)\rho + (1-\rho)} \end{aligned}$$

Now we aim to express $U = 2\int_0^1 \int_0^y (r(y)\text{-}r(x))dxdy$ in the form of $f(.)$. First, let

$$x = \rho(1 - f(t_g)) + (1-\rho)(1-t_g)$$

$$y = \rho(1 - f(s_g)) + (1-\rho)(1-s_g)$$

It follows that

$$\frac{dx}{dt_g} = -(1-\rho) - \rho f'(t_g)$$

Since $f(0) = 0$ and $f(1) = 1$, we know

$$x = 1 \Leftrightarrow t_g = 0$$

$$x = 0 \Leftrightarrow t_g = 1.$$

It follows then

$$\begin{aligned}
U &= 2\int_0^1 \int_0^y (r(y)\text{-}r(x))dxdy \\
&= 2\int_0^1 \int_0^s \left(\frac{f'(s)\rho}{f'(s)\rho + (1-\rho)} - \frac{f'(t)\rho}{f'(t)\rho + (1-\rho)}\right)[(1-\rho) + \rho f'(t)][(1-\rho) + \rho f'(s)]dtds \\
&= 2\int_0^1 f'(s)\rho \int_s^1 [(1-\rho) + \rho f'(t)]dtds - 2\int_0^1 [(1-\rho) + \rho f'(s)] \int_s^1 f'(t)\rho dtds \\
&= 2\int_0^1 f'(s)\rho[(1-\rho)(1-s) + \rho(1 - f(s))]ds - 2\int_0^1 [(1-\rho) + \rho f'(s)]ds \\
&= 2\rho(1-\rho)\int_0^1 [f'(s) - sf'(s) - 1 + f(s)]ds
\end{aligned}$$

In addition, because $\int_0^1 f'(s)ds = 1$ and $\int_0^1 sf'(s)ds = sf(s)|_0^1 - \int_0^1 f(s)ds$, the above equation can be simplified as

$$U = 2\rho(1-\rho)[2\int_0^1 f(s)ds - 1] = 2\rho(1-\rho)(2AUC_R - 1)$$

A.2. Connection between U-Index and Area Under the Lorenze Curve

We first show the connection between area under the ROC curve (AUC_R) and the area under the Lorenze Curve (AUC_L).

$$\begin{aligned}
AUC_L &= \frac{1}{\rho}\int_0^1\int_0^y r(x)dxdy \\
&= \int_0^1 [1-\rho+\rho f'(s)]ds\int_s^1 f'(t)dt \\
&= \int_0^1 [1- f(s)][1-\rho+\rho f'(s)]ds \\
&= (1-\rho)(1-\int_0^1 f(s)ds)+\rho\int_0^1 f'(s)ds-\rho\int_0^1 f(s)f'(s)ds \\
&= (1-\rho)(1-AUC_R)+\rho-\rho\int_0^1 f(s)f'(s)ds
\end{aligned}$$

Since $\int_0^1 f(s)f'(s)ds = f(s)f(s)|_0^1 - \int_0^1 f'(s)f(s)ds$ and $f(1)=1$, $f(0)=0$, we have

$$AUC_L = (1-\rho)(1-AUC_R)+\frac{1}{2}\rho$$

Further from $U=2\rho(1-\rho)(2AUC_R-1)$, it follows

$$U = 2\rho(0.5-AUC_L)$$

A.3. Connection between U-Index and two-sample U-Statistics

The U-Index can be estimated as

$$U = 2\sum_{1\le k<k'\le K} p_k p_{k'} \psi(r_k, r_{k'}) = 2\sum_{1\le k<k'\le K} p_k p_{k'} (r_k - r_{k'})\,;$$

where p_k and r_k are calculated from $P(G_k \mid D)$ and $P(G_k \mid \bar{D})$. As a result, we write U-Index as:

$$\begin{aligned}
U &= 2\sum_{k=1}^{K}\sum_{k'=1}^{k} p_k p_{k'} (r_k - r_{k'}) \\
&= 2\sum_{k=1}^{K} P(g_k, D)\sum_{k'=1}^{k} P(g_{k'}) - 2\sum_{k=1}^{K} P(g_k)\sum_{k'=1}^{k} P(g_{k'}, D) \\
&= 2\sum_{k=1}^{K} \rho P(g_k \mid D)\sum_{k'=1}^{k}[\rho P(g_{k'} \mid D) + (1-\rho)P(g_{k'} \mid \bar{D})] - 2\sum_{k=1}^{K}[\rho P(g_k \mid D) + (1-\rho)P(g_k \mid \bar{D})]\sum_{k'=1}^{k} \rho P(g_{k'} \mid D) \\
&= 2\rho(1-\rho)[\sum_{k=1}^{K} P(g_k \mid D)\sum_{k'=1}^{k} P(g_{k'} \mid \bar{D}) - \sum_{k=1}^{K} P(g_k \mid \bar{D})\sum_{k'=1}^{k} P(g_{k'} \mid D)]
\end{aligned}$$

It follows:

$$U = 2\rho(1-\rho)[\sum_{k=1}^{K}\sum_{k'=1}^{K} P(g_k \mid D) P(g_{k'} \mid \bar{D})(I_{\{k>k'\}} - I_{\{k<k'\}})];$$

where $I_{\{.\}}$ is an indicator function. Further, based on estimator

$$P(g_k \mid D) = \frac{n_{g_k,D}}{n_D} \text{ and } P(g_k \mid \bar{D}) = \frac{n_{g_k,\bar{D}}}{n_{\bar{D}}}$$

we can show that the U-Index is equivalent to a two-sample U-Statistic.

$$\begin{aligned}
U &= 2\rho(1-\rho)\frac{1}{n_D n_{\bar{D}}}[\sum_{k=1}^{K}\sum_{k'=1}^{K} n_{g_k,D} n_{g_{k'},\bar{D}} (I_{\{k>k'\}} - I_{\{k<k'\}})] \\
&= 2\rho(1-\rho)\frac{1}{n_D n_{\bar{D}}}\sum_{i=1}^{n_D}\sum_{j=1}^{n_{\bar{D}}} \phi(G_i, G_j)
\end{aligned}$$

where G_i is the genotype of the *i*-th subject in diseased population; G_j is the genotype of the *j*-th subject in non-diseased population. The kernel function has the following form:

$$\phi(G_i,G_j)=\begin{cases}1 & r(G_i)>r(G_j)\\ 0 & r(G_i)=r(G_j)\\ -1 & r(G_i)<r(G_j)\end{cases}$$

Further denote $\theta=E(\phi(G_i,G_j))$ and $\theta_U=E(U)=2\rho(1-\rho)\theta$. We can estimate the variance of

$$\begin{aligned}Var(U-\theta_U)&=\frac{4\rho^2(1-\rho)^2}{n_D^2n_{\bar{D}}^2}Var[\sum_{i-1}^{nD}\sum_{j=1}^{n\bar{D}}(\phi(G_i-G_j)-\theta)]\\&=\frac{4\rho^2(1-\rho)^2}{n_D^2n_{\bar{D}}^2}[n_Dn_{\bar{D}}\tau_{1,1}+n_Dn_{\bar{D}}(n_{\bar{D}}-1)\tau_{1,0}+n_Dn_{\bar{D}}(n_D-1)\tau_{0,1}]\end{aligned};$$

where $\tau_{1,1}=Var(\phi(G_i,G_j))$, $\tau_{1,0}=cov(\phi(G_i,G_j),\phi(G_i,G_{j'}))$ and $\tau_{0,1}=cov(\phi(G_i,G_j),\phi(G_{i'},G_j))$.

To obtain the asymptotic distribution of U, we can use Hajek projection to project $U-\theta_U$ onto the space of the summation forms $\sum_{i=1}^{n}h(G_i)$ where the CLT can be applied. The Hajek projection $\tilde{U}$ of $U-\theta_U$ is,

$$\begin{aligned}\tilde{U}&=\sum_{i=1}^{nD}E(U-\theta_U\mid G_i)+\sum_{j=1}^{n\bar{D}}E(U-\theta_U\mid G_j)\\&=\frac{2\rho(1-\rho)}{n_D}\sum_{i=1}^{nD}h_{1,0}(G_i)+\frac{2\rho(1-\rho)}{n_{\bar{D}}}\sum_{j=1}^{n\bar{D}}h_{0,1}(G_j)\end{aligned};$$

where $h_{1,0}(G_i)=E(\phi(G_i,G_j)-\theta\mid G_i)$ and $h_{0,1}(G_j)=E(\phi(G_i,G_j)-\theta\mid G_j)$. We can then calculate the variance of $\tilde{U}$ as

$$Var(\tilde{U}) = \frac{4\rho^2(1-\rho)^2}{n_D} Var(h_{1,0}(G_i)) + \frac{4\rho^2(1-\rho)^2}{n_{\bar{D}}} Var(h_{0,1}(G_j))$$

$$= 4\rho^2(1-\rho)^2 [\frac{\tau_{1,0}}{n_D} + \frac{\tau_{0,1}}{n_{\bar{D}}}]$$

We can write $U-\theta_U$ as a summation of the projection term U and the remaining term $\bar{\mathrm{R}}$, i.e. $U-\theta_U = \tilde{U} + \bar{\mathrm{R}}$. The asymptotic normality of $U-\theta_U$ is then established by showing is $\tilde{U}$ asymptotically normal and $\bar{\mathrm{R}}$ is asymptotically negligible. Assuming $n = n_D + n_{\bar{D}}$ and $\frac{n_D}{n} \to \lambda$, we can apply CLT to $\tilde{U}$ and show that

$$\sqrt{n}\tilde{U} \sim N(0, 4\rho^2(1-\rho)^2[\frac{\tau_{1,0}}{\lambda} + \frac{\tau_{0,1}}{1-\lambda}])$$

With the fact that $E(\tilde{U}) = 0, E(\bar{\mathrm{R}}) = 0$ and $E(\tilde{U}\bar{\mathrm{R}}) = 0$, we know

$$E(n\tilde{\mathrm{R}}^2) = nVar(U-\theta) - nVar(\tilde{U}) \to 0$$

Thus, $\sqrt{n}\bar{\mathrm{R}} \xrightarrow{p} 0$. With Slusky theorem, it follows that

$$\sqrt{n}(U-\theta) \sim N(0, 4\rho^2(1-\rho)^2[\frac{\tau_{1,0}}{\lambda} + \frac{\tau_{0,1}}{1-\lambda}])$$