

In previous results, we fit the outcome variable with a linear mixed model

$$Y = S(\beta + b + \Gamma(\alpha'x)) + \epsilon$$

and treat the coefficient $z = \beta + b + \Gamma(\alpha'x)$ as a MVN, that is, $z|w \sim MVN(\beta + \Gamma(\alpha'x), D)$.

We may also calculate the Kullback-Leibler divergence by using the outcome variables directly, by assuming

$$\begin{aligned} Y &= X(\beta + \Gamma(\alpha'x)) + Zb + \epsilon \\ Y &\sim MVN(S(\beta + \Gamma(\alpha'x)), ZDZ') \end{aligned}$$

Let's see whether they can return similar results.

1 Kullback-Leibler divergence and Purity

To measure how much the differences are between the treatment group and the placebo group, we apply the Kullback-Leibler (KL) divergence, which measures how one probability distribution F_1 is different from another probability distribution F_2 .

$$D_{KL}(F_1||F_2) = \int_{-\infty}^{+\infty} f_1(x) \log\left(\frac{f_1(x)}{f_2(x)}\right) dx \quad (1)$$

where f_1 and f_2 denote the probability density functions (pdf) of F_1 and F_2 , separately. The larger the KL divergence between distributions is, the more "pure" the distributions are. Besides, $D_{KL}(F_1||F_2) \geq 0$. Similarly, the $D_{KL}(F_2||F_1)$ is also always larger than or equals to 0.

Based on the Kullback-Leibler divergence, we define the *purity*, which represent how much the differences between the treatment group distribution F_1 and the placebo group distribution F_2 . We define the purity function of the summation of two Kullback-Leibler divergence as

$$\begin{aligned} \text{purity} &= D_{KL}(F_1||F_2) + D_{KL}(F_2||F_1) \\ &= \int_{-\infty}^{+\infty} f_1(x) \log\left(\frac{f_1(x)}{f_2(x)}\right) dx + \int_{-\infty}^{+\infty} f_2(x) \log\left(\frac{f_2(x)}{f_1(x)}\right) dx \end{aligned} \quad (2)$$

where

$$f_1(x) \sim MVN(\mu_1, \Sigma_1)$$

$$f_2(x) \sim MVN(\mu_2, \Sigma_2)$$

Let's calculate the purity value by calculating $\int f_1 \log f_1$, $\int f_2 \log f_2$, $\int f_1 \log f_2$, and $\int f_2 \log f_1$.

Part $\int f_1 \log f_1$

$$\begin{aligned} \int f_1 \log f_1 &= E_1 \left\{ -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log(|\Sigma_1|) - \frac{1}{2} (\mathbf{x} - \mu_1)' (\Sigma_1)^{-1} (\mathbf{x} - \mu_1) \right\} \\ &= -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log(|\Sigma_1|) - \frac{1}{2} E_1[(\mathbf{x} - \mu_1)' (\Sigma_1)^{-1} (\mathbf{x} - \mu_1)] \end{aligned}$$

And

$$\begin{aligned}
E_1[(\mathbf{x} - \boldsymbol{\mu}_1)'(\boldsymbol{\Sigma}_1)^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)] &= E_1[\text{tr}((\mathbf{x} - \boldsymbol{\mu}_1)'(\boldsymbol{\Sigma}_1)^{-1}(\mathbf{x} - \boldsymbol{\mu}_1))] \\
&= E_1[\text{tr}((\boldsymbol{\Sigma}_1)^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)(\mathbf{x} - \boldsymbol{\mu}_1)')] \\
&= \text{tr}(E_1[(\boldsymbol{\Sigma}_1)^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)(\mathbf{x} - \boldsymbol{\mu}_1)']) \\
&= \text{tr}(\boldsymbol{\Sigma}_1^{-1}E_1[(\mathbf{x} - \boldsymbol{\mu}_1)(\mathbf{x} - \boldsymbol{\mu}_1)']) \\
&= \text{tr}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_1) = \text{tr}(\mathbf{I}_n) = n
\end{aligned}$$

Therefore,

$$\int f_1 \log f_1 = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log(|\boldsymbol{\Sigma}_1|) - \frac{n}{2} \quad (3)$$

Similarly,

$$\int f_2 \log f_2 = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log(|\boldsymbol{\Sigma}_2|) - \frac{n}{2} \quad (4)$$

Part $\int f_1 \log f_2$

$$\begin{aligned}
\int f_1 \log f_2 &= E_1\left(-\frac{n}{2} \log(2\pi) - \frac{1}{2} \log(|\boldsymbol{\Sigma}_2|) - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_2^{-1}(\mathbf{x} - \boldsymbol{\mu}_2)\right) \\
&= -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log(|\boldsymbol{\Sigma}_2|) - \frac{1}{2} E_1[(\mathbf{x} - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_2^{-1}(\mathbf{x} - \boldsymbol{\mu}_2)]
\end{aligned}$$

And

$$\begin{aligned}
&E_1[(\mathbf{x} - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_2^{-1}(\mathbf{x} - \boldsymbol{\mu}_2)] \\
&= E_1[(\mathbf{x} - \boldsymbol{\mu}_1 + \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_2^{-1}(\mathbf{x} - \boldsymbol{\mu}_1 + \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)] \\
&= E_1[(\mathbf{x} - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}_2^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_2^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) \\
&\quad + (\mathbf{x} - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)] \\
&= E_1[(\mathbf{x} - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}_2^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)] + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_2^{-1} E_1(\mathbf{x} - \boldsymbol{\mu}_1) + \\
&\quad E_1(\mathbf{x} - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\
&= E_1[(\mathbf{x} - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}_2^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)] + 0 + 0 + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\
&= E_1[\text{tr}(\mathbf{x} - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}_2^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)] + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\
&= E_1[\text{tr}(\boldsymbol{\Sigma}_2^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)'(\mathbf{x} - \boldsymbol{\mu}_1))] + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\
&= \text{tr}(E_1[\boldsymbol{\Sigma}_2^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)'(\mathbf{x} - \boldsymbol{\mu}_1)]) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\
&= \text{tr}(\boldsymbol{\Sigma}_2^{-1}E_1[(\mathbf{x} - \boldsymbol{\mu}_1)'(\mathbf{x} - \boldsymbol{\mu}_1)]) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\
&= \text{tr}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)
\end{aligned}$$

Therefore,

$$\int f_1 \log f_2 = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log(|\boldsymbol{\Sigma}_2|) - \frac{1}{2} \{ \text{tr}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \} \quad (5)$$

Similarly,

$$\int f_2 \log f_1 = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log(|\boldsymbol{\Sigma}_1|) - \frac{1}{2} \{ \text{tr}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \} \quad (6)$$

Then the purity is

$$\begin{aligned}
& \int f_1 \log f_1 + \int f_2 \log f_2 - \int f_2 \log f_1 - \int f_1 \log f_2 \\
&= -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log(|\Sigma_1|) - \frac{n}{2} \\
& -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log(|\Sigma_2|) - \frac{n}{2} \\
& -(-\frac{n}{2} \log(2\pi) - \frac{1}{2} \log(|\Sigma_2|) - \frac{1}{2} \{tr(\Sigma_2^{-1} \Sigma_1) + (\mu_1 - \mu_2)' \Sigma_2^{-1} (\mu_1 - \mu_2)\}) \\
& -(-\frac{n}{2} \log(2\pi) - \frac{1}{2} \log(|\Sigma_1|) - \frac{1}{2} \{tr(\Sigma_1^{-1} \Sigma_2) + (\mu_1 - \mu_2)' \Sigma_1^{-1} (\mu_1 - \mu_2)\}) \\
&= -n + \frac{1}{2} tr(\Sigma_1^{-1} \Sigma_2) + \frac{1}{2} tr(\Sigma_2^{-1} \Sigma_1) \\
& + \frac{1}{2} [(\mu_1 - \mu_2)' \Sigma_1^{-1} (\mu_1 - \mu_2)] + \frac{1}{2} [(\mu_1 - \mu_2)' \Sigma_2^{-1} (\mu_1 - \mu_2)]
\end{aligned} \tag{7}$$

Then the purity is defined as $-n + \frac{1}{2} tr(\Sigma_1^{-1} \Sigma_2) + \frac{1}{2} tr(\Sigma_2^{-1} \Sigma_1) + \frac{1}{2} [(\mu_1 - \mu_2)' \Sigma_1^{-1} (\mu_1 - \mu_2)] + \frac{1}{2} [(\mu_1 - \mu_2)' \Sigma_2^{-1} (\mu_1 - \mu_2)]$ for two normal distributions f_1, f_2 with mean μ_1, μ_2 respectively and covariance matrix Σ_1, Σ_2 respectively.

Back to our model, when we fit the coefficients of the LME, i.e.,

$$z = \beta + b + \Gamma(\alpha'x)$$

as multivariate normal distributions and plug in equation (7), we can get our purity function:

$$Purity(\alpha) = A_0 + A_1 \mu'_x \alpha + \frac{A_2}{2} [\alpha' \Sigma_x \alpha + \alpha' \mu_x \mu'_x \alpha] \tag{8}$$

where

$$\begin{aligned}
A_0 &= -q + \frac{1}{2} tr(D_2^{-1} D_1) + \frac{1}{2} tr(D_1^{-1} D_2) + \frac{1}{2} (\beta_1 - \beta_2)' (D_1^{-1} + D_2^{-1}) (\beta_1 - \beta_2) \\
A_1 &= (\beta_1 - \beta_2)' (D_1^{-1} + D_2^{-1}) (\Gamma_1 - \Gamma_2) \\
A_2 &= (\Gamma_1 - \Gamma_2)' (D_1^{-1} + D_2^{-1}) ((\Gamma_1 - \Gamma_2))
\end{aligned}$$

When we fit the outcome as normal distribution and plug in the equation (7), we can simply replace the β in equation (8) as $S\beta$; replace Γ as $S\Gamma$; replace D as ZDZ'

Replace μ_1, μ_2 with $X(\beta_1 + \Gamma_1(\alpha'x))$, and $X(\beta_2 + \Gamma_2(\alpha'x))$. Replace D_1, D_2 with ZD_1Z', ZD_2Z' . Then the purity function is

$$Purity(\alpha) = B_0 + B_1 \mu'_x \alpha + \frac{B_2}{2} [\alpha' \Sigma_x \alpha + \alpha' \mu_x \mu'_x \alpha] \tag{9}$$

where

$$\begin{aligned}
B_0 &= -q + \frac{1}{2} tr((ZD_2Z')^{-1} (ZD_1Z')) + \frac{1}{2} tr((ZD_1Z')^{-1} (ZD_2Z')) \\
& + \frac{1}{2} (S\beta_1 - S\beta_2)' ((ZD_1Z')^{-1} + (ZD_2Z')^{-1}) (S\beta_1 - S\beta_2) \\
B_1 &= (S\beta_1 - S\beta_2)' ((ZD_1Z')^{-1} + (ZD_2Z')^{-1}) (S\Gamma_1 - S\Gamma_2) \\
B_2 &= (S\Gamma_1 - S\Gamma_2)' ((ZD_1Z')^{-1} + (ZD_2Z')^{-1}) ((S\Gamma_1 - S\Gamma_2))
\end{aligned}$$