Outline 3

February 20, 2019

0.1 Step1:

I was confused about the dimensions of the following equation. I just wanted to add the subscript i in the equation and specify the euqation for each subject.

To calculate the max purity function, first, fit the linear mixed model for the outcome $\boldsymbol{y}_i$and time $\boldsymbol{X}_i$, with baseline covariates $\boldsymbol{x}_i$:

$$\boldsymbol{y}_i = \boldsymbol{X}_i(\boldsymbol{\beta} + \boldsymbol{b}_i + \boldsymbol{\Gamma}(\boldsymbol{\alpha}'\boldsymbol{x}_i)) + \boldsymbol{\epsilon}_i.$$

where,

- $\boldsymbol{y}_i$ is the vector of outcome for the ith subject, i.e., the dimension of $\boldsymbol{y}_i$ is $(n_i, 1)$. n_i is the number of observations for subject i.
- $\boldsymbol{X}_i$ is the covariance matrix for the ith subject. The dimension is n_i, p. P is the number of covariates. Here in our example $p = 3$, which is for $(1, t, t^2)$.
- $\boldsymbol{\beta}$ is the coefficient vector for the fixed effects of $\boldsymbol{X}_i$. The dimension is $(p, 1)$.
- $\boldsymbol{x}_i$ is the vector of the baseline covariates for the subject. The dimension is $(q, 1)$.
- $\boldsymbol{b}_i$ is the vector of random effects. The dimension is $(p, 1)$.
- $\boldsymbol{\Gamma}$ is the vector of fixed effects of the baseline covariates. Dimension is $(p, 1)$.

- $w_i = \alpha' \boldsymbol{x}_i$ is the combination of the input baseline covariates. w_i is a scalar.

We can define the covariance matrix of $\boldsymbol{X}_i$ as $\boldsymbol{z}_i$. The $\boldsymbol{z}_i$ contains both fixed effects and random effects.

$$\boldsymbol{z}_i = \boldsymbol{\beta} + \boldsymbol{b}_i + \boldsymbol{\Gamma} w_i$$

We can also write the above equation in the matrix version:

$$Y_i = \begin{bmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ \dots & \dots & \dots \\ 1 & t_{n_i} & t_{n_i}^2 \end{bmatrix} \left[\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} + \begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \end{pmatrix} w_i \right] + \boldsymbol{\epsilon}_i$$

For different subjects, they have the same $\boldsymbol{\beta}$ vector and the same $\boldsymbol{\Gamma}$ vector. Their random effect vector $\boldsymbol{b}_i$ can be different.
Sorry I am still a little confused here. How could we combine the $\boldsymbol{Y}_i$, $\boldsymbol{X}_i$ and the matrix of coefficients of each subject together to get an overall equation, like $\boldsymbol{Y} = \boldsymbol{X}$ times some matrix?

0.2 Step 2:

Estimate the distribution of $\boldsymbol{z}_i$ for drug group and placebo, separately.

$$\begin{aligned} f(\boldsymbol{z}_i) &= \int_{w_i} f(\boldsymbol{z}_i, w_i) dw_i \\ &= \int_{w_i} f(\boldsymbol{z}_i | w_i) g(w_i) dw_i \end{aligned}$$

where,

- the conditional distribution $f(\boldsymbol{z}_i | w_i) \sim MVN(\boldsymbol{\beta} + \boldsymbol{\Gamma} w_i, \boldsymbol{D})$
- $g(w_i)$ is the distribution of the covariates combination $\boldsymbol{\alpha}' \boldsymbol{x}_i$. For example, if the covariates combination only contains "sex", which is binary, then the integral becomes summation.
- $\boldsymbol{D}_i$ is the covariance matrix of random effects $\boldsymbol{b}_i$.

We could then estimate the $f(\cdot)$ for drug group and placebo group separately, i.e. $f_1(\boldsymbol{z}_i)$ and $f_2(\boldsymbol{z}_i)$.

0.3 Step 3:

The purity should be a function of w_i. The integral of z_i should be calculated first. Then we can define the purity as,

$$\begin{aligned} P_{w_i} &= \int_{z_i} \frac{[f_1(z_i|w_i) - f_2(z_i|w_i)]^2}{[f_1(z_i|w_i) + f_2(z_i|w_i)]^2}(f_1(z_i|w_i) + f_2(z_i|w_i))dz_i \\ &= \int_{z_i} \frac{[f_1(z_i|w_i) - f_2(z_i|w_i)]^2}{f_1(z_i|w_i) + f_2(z_i|w_i)}dz_i \end{aligned}$$

where

- the integral of with high dimension, which is equals to the dimension of $\boldsymbol{z}_i$.
- the $f_1(z_i|w_i)$ and $f_2(z_i|w_i)$ are pdf of multivariate normal distributions.
- $f_1(\boldsymbol{z}_i|w_i) = (2\pi)^{-\frac{p}{2}}|\boldsymbol{D}|^{-1/2}\exp(-\frac{1}{2}(\boldsymbol{z}_i - \hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\Gamma}}_1 w_i)^T D^{-1}(\boldsymbol{z}_i - \hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\Gamma}}_1 w_i))$
- $f_2(\boldsymbol{z}_i|w_i) = (2\pi)^{-\frac{p}{2}}|\boldsymbol{D}|^{-1/2}\exp(-\frac{1}{2}(\boldsymbol{z}_i - \hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\Gamma}}_1 w_i)^T D^{-1}(\boldsymbol{z}_i - \hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\Gamma}}_1 w_i))$

It can be also approximated as the summation form:

$$\sum_{i=1}^{n}\Big[\frac{f_1(\boldsymbol{z}_i|w_i) - f_2(\boldsymbol{z}_i|w_i)}{f_1(\boldsymbol{z}_i|w_i) + f_2(\boldsymbol{z}_i|w_i)}\Big]^2$$

The summation is also high dimensional, with the same dimension as $\boldsymbol{z}_i$.
This is one purity value based on one combination of baseline covariates for subject i.
Then we can get the subject purity, i.e., the purity for subject $\boldsymbol{w}_i$. We need to calculate the integral or summation of $\boldsymbol{w}_i$ next to get an overall purity for the data.
We would like to find the $\boldsymbol{w}$ or $\boldsymbol{\alpha}$ that max the $\boldsymbol{P}_{\boldsymbol{\alpha}'\boldsymbol{x}}$, i.e.,

$$\begin{aligned} \hat{\boldsymbol{\alpha}} &= \arg\max_{\boldsymbol{\alpha}} P_{\boldsymbol{\alpha}} \\ &= \arg\max_{\boldsymbol{\alpha}} \int_{\boldsymbol{\alpha}'\boldsymbol{x}} P_{\boldsymbol{\alpha}'\boldsymbol{x}} d\boldsymbol{\alpha}'\boldsymbol{x} \\ &= \arg\max_{\boldsymbol{\alpha}} \int_{\boldsymbol{\alpha}'\boldsymbol{x}} \frac{(f_1(\boldsymbol{z}|\boldsymbol{\alpha}'\boldsymbol{x}) - f_2(\boldsymbol{z}|\boldsymbol{\alpha}'\boldsymbol{x}))^2}{1 f_1(\boldsymbol{z}|\boldsymbol{\alpha}'\boldsymbol{x}) + f_2(\boldsymbol{z}|\boldsymbol{\alpha}'\boldsymbol{x})} d\boldsymbol{\alpha}'\boldsymbol{x} \end{aligned}$$