Discovering Linear Biosignatures for Treatment Response: A Convexity-Based Clustering Approach

2019-07-31

This is an outline about how could we paragraph the sections

1 Abstract

2 Introduction

1. A brief introduction of the psychiatric diseases and their difficulty to classify subjects who responded to the treatment. i.e. in a randomized clinical trial comparing treatments to placebo for mental illnesses, there often have subjects with different groups that have similar outcomes.

2. Review current approaches to solve the above problem. Introduce convex-based clustering algorithm.

3. However, those methods focus on the outcomes while ignoring subjects' biosignatures. Introduce the new method, which considers the baseline covariates. Introduce our aims.

4. Mention a little bit of the example: the EMBARC study

5. Organizations of sections.

3 Example: EMBARC study

1. A brief introduction of the EMBARC data.

2. Draw the trajectory of outcome to show the difficulty in clustering treatment group and find the homogeneous performance

We have more analysis of the EMBARC data Section 6.

4 Method

4.1 Model

Clustering the raw data will often give results similar to clustering regression coefficients obtained using an orthogonal design matrix.

In our setting, we assume the outcomes are from a linear mixed model:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i(\boldsymbol{\beta}_i + \boldsymbol{b}_i) + \boldsymbol{\epsilon}, i \in \{1, 2\} \tag{1}$$

where,

- $\boldsymbol{\beta}_i$ is the vector of covariates for fixed effects of $\boldsymbol{X}$
- $\boldsymbol{b}_i$ is the vector of random effects
- $\boldsymbol{\Gamma}_i$ is the vector of fixed effects of the baseline covariates.

Define the covariate matrix of $\boldsymbol{X}$ as $\boldsymbol{z}$. The $\boldsymbol{z}$ contains both fixed effects and random effects.

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

For subjects in the treated group (call group 1) and placebo group (call group 2), if the treatment have effect, the distribution of regression coefficients $\boldsymbol{z}_i$ should be different. The larger the difference of $\boldsymbol{z}_i$ is, the larger the difference of the treatment effect between group 1 and group 2.

However, this method does not consider the baseline covariates, which may affect the outcome. Therefore, we made a new model based on the Eq 1:

$$\boldsymbol{Y} = \boldsymbol{X}(\boldsymbol{\beta} + \boldsymbol{b} + \boldsymbol{\Gamma}(W)) + \boldsymbol{\epsilon}, \tag{2}$$

where,

- $\boldsymbol{\beta}$ is the vector of covariates for fixed effects of $\boldsymbol{X}$
- $\boldsymbol{b}$ is the vector of random effects
- $\boldsymbol{\Gamma}$ is the vector of fixed effects of the baseline covariates.
- $W = \boldsymbol{\alpha}'\boldsymbol{x}$ is the combination of the input baseline covariates.

Define the covariate matrix of $\boldsymbol{X}$ as $\boldsymbol{z}$. The $\boldsymbol{z}$ contains both fixed effects and random effects.

$$\boldsymbol{z} = \boldsymbol{\beta} + \boldsymbol{b} + \boldsymbol{\Gamma}\boldsymbol{w}$$

That is, we have distributions for the mixed-effect model coefficients $\boldsymbol{z}$ given $w = \boldsymbol{\alpha}'\boldsymbol{x}$, where

$$\boldsymbol{z}|w \sim N(\boldsymbol{\beta}_j + \boldsymbol{\Gamma}_j w, \boldsymbol{D}_j),$$

for treatment $j = 1, 2$.

We are aiming to determine an $\boldsymbol{\alpha}$, which is a linear combination of baseline measures that maximizes the distance between $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$, and cluster outcome trajectories so that clusters are homogeneous with respect to different treatment groups.

4.2 The application of baseline measures: purity

a) Kullback-Leibler divergence

In statistics, the Kullback-Leibler (KL) divergence is a measure of how one probability distribution F_1 is different from a sceond reference probability distribution F_2. We may apply the KL divergence idea in the clustering problem, since the larger the KL divergence between distributions, the more pure the groups/clusters are. For distributions F_1 and F_2 of a continuous random variable, the KL divergence is defined as:

$$D_{KL}(F_1||F_2) = \int_{-\infty}^{+\infty} f_1(x) log(\frac{f_1(x)}{f_2(x)}) dx \tag{3}$$

where f_1 and f_2 denote the probability density of F_1 and F_2.

The $D_{KL}(F_1||F_2)$ is always bigger or equal to than 0. Similarly, the $D_{KL}(F_2||F_1)$ is also always bigger or equal to than 0.

b). The definition of purity

We define a purity function, which is associated with the distance between distribution F_1 and F_2. The higher the purity is, the larger the distance between F_1 and F_2.

Define the **purity function** w.r.t the combination of covariates w as:

conditional purity

$$\begin{aligned} g(w) =& D_{KL}(F_1||F_2) + D_{KL}(F_2||F_1) \\ =& \int log(f_1(\boldsymbol{z}|\boldsymbol{w})) f_1(\boldsymbol{z}|\boldsymbol{w}) dz - \int log(f_2(\boldsymbol{z}|\boldsymbol{w})) f_1(\boldsymbol{z}|\boldsymbol{w}) dz \\ &+ \int log(f_2(\boldsymbol{z}|\boldsymbol{w})) f_2(\boldsymbol{z}|\boldsymbol{w}) dz - \int log(f_1(\boldsymbol{z}|\boldsymbol{w})) f_2(\boldsymbol{z}|\boldsymbol{w}) dz \end{aligned} \tag{4}$$

where,

- $f_1(\boldsymbol{z}|\boldsymbol{w}) = \frac{1}{\sqrt{((2\pi)^p|\boldsymbol{D}_1|)}} exp(-\frac{1}{2}(\boldsymbol{z} - \boldsymbol{\mu}_1)' \boldsymbol{D}_{\boldsymbol{1}}^{-1} (\boldsymbol{z} - \boldsymbol{\mu}_1))$
- $f_2(\boldsymbol{z}|\boldsymbol{w}) = \frac{1}{\sqrt{((2\pi)^p|\boldsymbol{D}_2|)}} exp(-\frac{1}{2}(\boldsymbol{z} - \boldsymbol{\mu}_2)' \boldsymbol{D}_2^{-1} (\boldsymbol{z} - \boldsymbol{\mu}_2))$
- $\boldsymbol{\mu}_1 = \boldsymbol{\beta}_1 + \boldsymbol{\Gamma}_1 w, \boldsymbol{\mu}_2 = \boldsymbol{\beta}_2 + \boldsymbol{\Gamma}_2 w$

The **purity function** regards to the whole dataset is:

$$\text{purity}(\boldsymbol{\alpha}) = \int_{\boldsymbol{w}} g(w) dw = \int_{\boldsymbol{x}} g(\boldsymbol{\alpha}'\boldsymbol{x}) \tag{5}$$

which can be estimated as $\sum_{\boldsymbol{x}} g(\boldsymbol{\alpha}'\boldsymbol{x})$

We aimed at determining the $\boldsymbol{alpha}$ that can maximize the purity($\boldsymbol{\alpha}$) function. We may apply **Grid Search** method to achieve that.

```
Algorithm Optimization of Purity by Grid Search
1: ~~Select baseline covariates $x$, with dimension $p$.~~    before algorithm
step 0
2: Initial $\boldsymbol{\alpha} = [\alpha_1, ..., \alpha_p]$
3: Calculate $W = \boldsymbol{\alpha}'\boldsymbol{x}$
4: For $\forall w \in W$, fit the models $\boldsymbol{Y} = \boldsymbol{X}(\boldsymbol{\beta} + \boldsymbol{b} + \boldsymbol{\Gamma}(w)) + \boldsymbol{\epsilon}$ from data in group 1 and group
   2, separately.
5: Estimate $\boldsymbol{\beta}_j$, $\boldsymbol{\Gamma}_j$ and $\boldsymbol{D}_j$, $j = \{1, 2\}$. Distribution 1 $\sim N(\beta_1 + \Gamma_1 w, D_1)$; Distribution 2
   $\sim N(\beta_2 + \Gamma_2 w, D_2)$
6: Generate a large sample $\boldsymbol{Z} \sim N(\boldsymbol{0}, I)$, transform $\boldsymbol{z}$ to distribution 1, as $z_1$, transform $\boldsymbol{z}$
   to distribution 2, as $z_2$
   summation, separate formula before the algorithm
7: Calculate $D = log_1(\frac{f_1}{f_2}) + log_2(\frac{f_1}{f_2})$.    7,8 single step: estimated the expectation of kl, refer the eq number
8: Calculate the mean value of $D$, based on different $w \in W$
9: Change initial $\boldsymbol{\alpha}$, repeat step 3-8. Find the $\boldsymbol{\alpha}$ that maximizes $\bar{D}$ =0
```

We may derive the solution of max($\boldsymbol{\alpha}$) from **Eq 4** to avoid long running time by grid search. As we have known:

$$\frac{\partial(purity(\boldsymbol{\alpha}))}{\partial \boldsymbol{\alpha}} \equiv 0 \tag{6}$$

The solution of Eq (6) is the $\boldsymbol{\alpha}$ that maximizes the purity function, which is:

$$\boldsymbol{\alpha} = -2B^{-1}A' \tag{7}$$

where

- $A = \sum_i (\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2)'(\boldsymbol{D}_1^{-1} + \boldsymbol{D}_2^{-1})(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)\boldsymbol{x}_i'$
- $B = \sum_i \boldsymbol{x}_i\big((\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2)\big)'(\boldsymbol{D}_1^{-1} + \boldsymbol{D}_2^{-1})\big((\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2)\big)\boldsymbol{x}_i'$

The Summary of the algorithm:

```
Algorithm Optimization of Purity by Derivation of KL-Function
no
1: Select baseline covariates $\boldsymbol{x}$, with dimension $p$.
0
2: Initial $\boldsymbol{\alpha} = [\alpha_1, ..., \alpha_p]$
3: Calculate $W = \boldsymbol{\alpha}'\boldsymbol{x}$
For the sepcific w from 3
4: For $\forall w \in W$, fit the models $\boldsymbol{Y} = \boldsymbol{X}(\boldsymbol{\beta} + \boldsymbol{b} + \boldsymbol{\Gamma}(w)) + \boldsymbol{\epsilon}$ from data in group 1 and group
   2, separately.    heads, current alpha
5: Estimate $\boldsymbol{\beta}_j$, $\boldsymbol{\Gamma}_j$ and $\boldsymbol{D}_j$, $j = \{1, 2\}$.
6: Solve $\boldsymbol{\alpha} = -2B^{-1}A'$, where $A = \sum_i (\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2)'(\boldsymbol{D}_1^{-1} + \boldsymbol{D}_2^{-1})(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)\boldsymbol{x}_i'$, $B = \sum_i \boldsymbol{x}_i\big((\boldsymbol{\Gamma}_1 -
   \boldsymbol{\Gamma}_2)\big)'(\boldsymbol{D}_1^{-1} + \boldsymbol{D}_2^{-1})\big((\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2)\big)$
7: Wrap the whole procedures as a fuction and use *optim* to find the maximum. =0
```

NR algorithm, details of the optim function.

4.3 Convexity-based clustering, our model

The purpose by conducting convexity-based clustering is to find a partition of the dataset that maximize the homogeneous.

Summary of the algorithm

Algorithm Steps of the Convexity-Based Clustering

```
1: Find the $\boldsymbol{\alpha}$ that maximizes the purity function
2: Initialize a partition $B_1, B_2.., B_k$ (k clusters)
3: Calculate the support points $h_j = \frac{\pi_2 P_2(Bj)}{P(B_j)}$, $P(B_j) = \pi_1 P_2(Bj) + \pi_2 P_2(Bj)$
4: Determine a minimum support plane partition $D_j = \{||\lambda - h_j|| < ||\lambda - h_i||, i \neq j\}$,
   $\lambda(x) = \frac{\pi_2 f_2(x)}{f(x)}$ is the posterior probability that an observation $x$ belongs to population
   II.
5: Update the partition by $B_j \leftarrow \lambda^{-1}(D_j)$
6: Repeat 3-5 until the convergence criterion is met =0
```

5 Simulation

We conducted a simulation to illustrate that the α we get from the derivated formula can achieve the maximum purity value.

1. Data generation

   Settings:

   - Two groups: treatment and placebo, each group has 100 subjects. Each subject has 7 measure times.
   - Two baseline covariates, $\boldsymbol{x} = [x_1, x_2]$, $x_1, x_2 \sim N(0,1)$
   - Set $\beta_i, \Gamma_i, D_i, i \in \{1,2\}$; set $\boldsymbol{\alpha}, s.t. W = \boldsymbol{\alpha}'\boldsymbol{x}$
   - $\epsilon \sim N(0,1)$
   - Generate outcomes of group 1 and group 2: $Y_i = X_i(\beta_i + \Gamma_i W_i) + \epsilon_i$

2. Find $\boldsymbol{\alpha}$ Apply the above algorithm to find the $\boldsymbol{\alpha}$ 3. Results

   - Plot: α v.s. purity
   - Cluster table: the one with α has better performance than the one that doesn't have.

4. Conclusion: the α that maximizes the purity matches the true α. We may use the derivated approach to find the α in the model.

6 Application on the EMBARC study

1. A brief introduction of the EMBARC data.

2. Demographics, demographics table

3. The choices of two baseline variates.

4. Results:

- clustering table
- trajectory plots
- boundary plots
- ellipse plots

5. Comparison of results without consideration of the linear combination of baseline covariates.

7 Discussion and conclusion

8 Citations

9 Supplementary

9.1 Derivation of $\boldsymbol{\alpha}$

We can separate Equation(2) into four parts: $\int f_1 log f_1$, $\int f_2 log f_2$, $\int f_1 log f_2$, and $\int f_2 log f_1$.

- For $\int f_1 log f_1$:

$$\begin{aligned}\int f_1 log f_1 =& E_1(-\frac{p}{2}log(2\pi) - \frac{1}{2}log(|\boldsymbol{D}_1|) - \frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu}_1)'\boldsymbol{D}_1^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_1))\\ =& -\frac{p}{2}log(2\pi) - \frac{1}{2}log(|\boldsymbol{D}_1|) - \frac{1}{2}E_1[(\boldsymbol{z}-\boldsymbol{\mu}_1)'\boldsymbol{D}_1^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_1)]\end{aligned}$$

And

$$\begin{aligned}E_1[(\boldsymbol{z}-\boldsymbol{\mu}_1)'\boldsymbol{D}_1^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_1)] =& E_1[tr((\boldsymbol{z}-\boldsymbol{\mu}_1)'\boldsymbol{D}_1^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_1))]\\ =& E_1[tr(\boldsymbol{D}_1^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_1)'(\boldsymbol{z}-\boldsymbol{\mu}_1))]\\ =& tr(E_1[\boldsymbol{D}_1^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_1)'(\boldsymbol{z}-\boldsymbol{\mu}_1)])\\ =& tr(\boldsymbol{D}_1^{-1}E_1[(\boldsymbol{z}-\boldsymbol{\mu}_1)'(\boldsymbol{z}-\boldsymbol{\mu}_1)])\\ =& tr(\boldsymbol{D}_1^{-1}\boldsymbol{D}_1) = tr(\boldsymbol{I}_p) = p\end{aligned}$$

Therefore,

$$\int f_1 log f_1 = -\frac{p}{2}log(2\pi) - \frac{1}{2}log(|\boldsymbol{D}_1|) - \frac{p}{2} \tag{8}$$

Similarly,

$$\int f_2 log f_2 = -\frac{p}{2}log(2\pi) - \frac{1}{2}log(|\boldsymbol{D}_2|) - \frac{p}{2} \tag{9}$$

- For $\int f_1 log f_2$

$$\begin{aligned}\int f_1 log f_2 =& E_1(-\frac{p}{2}log(2\pi) - \frac{1}{2}log(|\boldsymbol{D}_2|) - \frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu}_2)'\boldsymbol{D}_2^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_2))\\ =& -\frac{p}{2}log(2\pi) - \frac{1}{2}log(|\boldsymbol{D}_2|) - \frac{1}{2}E_1[(\boldsymbol{z}-\boldsymbol{\mu}_2)'\boldsymbol{D}_2^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_2)]\end{aligned}$$

And

$$\begin{aligned}E_1[(\boldsymbol{z}-\boldsymbol{\mu}_2)'\boldsymbol{D}_2^{-1}(z-\boldsymbol{\mu}_2)] =& E_1[(\boldsymbol{z}-\boldsymbol{\mu}_1+\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{D}_2^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_1+\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)]\\ =& E_1[(\boldsymbol{z}-\boldsymbol{\mu}_1)'\boldsymbol{D}_2^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_1)+(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{D}_2^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_1)\\ &+(\boldsymbol{z}-\boldsymbol{\mu}_1)\boldsymbol{D}_2^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)+(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{D}_2^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)]\\ =& E_1[(\boldsymbol{z}-\boldsymbol{\mu}_1)'\boldsymbol{D}_2^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_1)]+(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{D}_2^{-1}E_1(\boldsymbol{z}-\boldsymbol{\mu}_1)+\\ &E_1(\boldsymbol{z}-\boldsymbol{\mu}_1)')D_2^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)+(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{D}_2^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\\ =& E_1[(\boldsymbol{z}-\boldsymbol{\mu}_1)'\boldsymbol{D}_2^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_1)]+0+0+(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{D}_2^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\\ =& E_1[tr(\boldsymbol{z}-\boldsymbol{\mu}_1)'\boldsymbol{D}_2^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_1))]+(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{D}_2^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\\ =& E_1[tr(\boldsymbol{D}_2^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_1)'(\boldsymbol{z}-\boldsymbol{\mu}_1))]+(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{D}_2^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\\ =& tr(E_1[\boldsymbol{D}_2^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_1)'(\boldsymbol{z}-\boldsymbol{\mu}_1)])+(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{D}_2^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\\ =& tr(\boldsymbol{D}_2^{-1}E_1[(z-\boldsymbol{\mu}_1)'(\boldsymbol{z}-\boldsymbol{\mu}_1)])+(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{D}_2^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\\ =& tr(\boldsymbol{D}_2^{-1}\boldsymbol{D}_1)+(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{D}_2^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\end{aligned}$$

Therefore,

$$\int f_1 log f_2 = -\frac{p}{2}log(2\pi) - \frac{1}{2}log(|\boldsymbol{D}_2|) - \frac{1}{2}\big(tr(\boldsymbol{D}_2^{-1}\boldsymbol{D}_1)+(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{D}_2^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\big) \tag{10}$$

Similarly,

$$\int f_2 log f_1 = -\frac{p}{2}log(2\pi) - \frac{1}{2}log(|\boldsymbol{D}_1|) - \frac{1}{2}\big(tr(\boldsymbol{D}_1^{-1}\boldsymbol{D}_2)+(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{D}_1^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\big) \tag{11}$$

Therefore, the equation (1) is:

$$(3) = (5) - (7) + (6) - (8)$$

That is,

$$\begin{aligned}&\int log(f_1)f_1 - \int log(f_2)f_1 + \int log(f_2)f_2 - \int log(f_1)f_2\\ &= \big(-\frac{p}{2}log(2\pi) - \frac{1}{2}log(|\boldsymbol{D}_1|) - \frac{p}{2}\big)\\ &- \big(-\frac{p}{2}log(2\pi) - \frac{1}{2}log(|\boldsymbol{D}_2|) - \frac{1}{2}\big(tr(\boldsymbol{D}_2^{-1}\boldsymbol{D}_1)+(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{D}_2^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\big)\big)\\ &+ \big(-\frac{p}{2}log(2\pi) - \frac{1}{2}log(|\boldsymbol{D}_2|) - \frac{p}{2}\big)\\ &- \big(-\frac{p}{2}log(2\pi) - \frac{1}{2}log(|\boldsymbol{D}_1|) - \frac{1}{2}\big(tr(\boldsymbol{D}_1^{-1}\boldsymbol{D}_2)+(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{D}_1^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\big)\big)\\ &= -p + \frac{1}{2}tr(\boldsymbol{D}_2^{-1}\boldsymbol{D}_1) + \frac{1}{2}tr(\boldsymbol{D}_1^{-1}\boldsymbol{D}_2) + \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'(\boldsymbol{D}_1^{-1}+\boldsymbol{D}_2^{-1})(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\end{aligned}$$

where $\boldsymbol{\mu}_1 = \boldsymbol{\beta}_1 + \boldsymbol{\Gamma}_1\boldsymbol{\alpha}'\boldsymbol{x}$, $\boldsymbol{\mu}_2 = \boldsymbol{\beta}_2 + \boldsymbol{\Gamma}_2\boldsymbol{\alpha}'\boldsymbol{x}$.

Therefore,

$$
\begin{aligned}
&(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'(\boldsymbol{D}_1^{-1} + \boldsymbol{D}_2^{-1})(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\\
&= \big(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 + (\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2)\boldsymbol{\alpha}'\boldsymbol{x}\big)'(\boldsymbol{D}_1^{-1} + \boldsymbol{D}_2^{-1})\big(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 + (\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2)\boldsymbol{\alpha}'\boldsymbol{x}\big)\\
&= (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)'(\boldsymbol{D}_1^{-1} + \boldsymbol{D}_2^{-1})(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) + (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)'(\boldsymbol{D}_1^{-1} + \boldsymbol{D}_2^{-1})(\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2)\boldsymbol{\alpha}'\boldsymbol{x}\\
&+ \big(\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2)\boldsymbol{\alpha}'\boldsymbol{x}\big)'(\boldsymbol{D}_1^{-1} + \boldsymbol{D}_2^{-1})\big(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\big) + \big((\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2)\boldsymbol{\alpha}'\boldsymbol{x}\big)'(\boldsymbol{D}_1^{-1} + \boldsymbol{D}_2^{-1})\big((\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2)\boldsymbol{\alpha}'\boldsymbol{x}\big)\\
&= (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)'(\boldsymbol{D}_1^{-1} + \boldsymbol{D}_2^{-1})(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)\\
&+ \big[(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)'(\boldsymbol{D}_1^{-1} + \boldsymbol{D}_2^{-1})(\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2) + (\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2)'(\boldsymbol{D}_1^{-1} + \boldsymbol{D}_2^{-1})(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)\big]\boldsymbol{x}'\boldsymbol{\alpha}\\
&+ \boldsymbol{\alpha}'\boldsymbol{x}\big((\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2)\big)'(\boldsymbol{D}_1^{-1} + \boldsymbol{D}_2^{-1})\big((\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2)\big)\boldsymbol{x}'\boldsymbol{\alpha}
\end{aligned}
$$

Determine $\boldsymbol{\alpha}$ to maximize the Equation(5)

$$
\begin{aligned}
\frac{\partial(purity(\boldsymbol{\alpha}))}{\partial\boldsymbol{\alpha}} = \frac{\partial(\sum_i g(\boldsymbol{\alpha}'\boldsymbol{x}_i))}{\partial\boldsymbol{\alpha}} &= \sum_i \frac{\partial(g(\boldsymbol{\alpha}'\boldsymbol{x}_i))}{\partial\boldsymbol{\alpha}}\\
&= \sum_i \big[(\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2)'(\boldsymbol{D}_1^{-1} + \boldsymbol{D}_2^{-1})(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)\boldsymbol{x}_i'\boldsymbol{\alpha}\\
&+ \sum_i \boldsymbol{\alpha}'\boldsymbol{x}_i\big((\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2)\big)'(\boldsymbol{D}_1^{-1} + \boldsymbol{D}_2^{-1})\big((\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2)\big)\boldsymbol{x}_i'\boldsymbol{\alpha}\big]'\\
&= \sum_i \big[(\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2)'(\boldsymbol{D}_1^{-1} + \boldsymbol{D}_2^{-1})(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)\boldsymbol{x}_i'\big]\\
&+ \sum_i \boldsymbol{\alpha}'\big[\boldsymbol{x}_i\big((\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2)\big)'(\boldsymbol{D}_1^{-1} + \boldsymbol{D}_2^{-1})\big((\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2)\big)\boldsymbol{x}_i'\big]\\
&+ \sum_i \boldsymbol{\alpha}'\big[\boldsymbol{x}_i\big((\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2)\big)'(\boldsymbol{D}_1^{-1} + \boldsymbol{D}_2^{-1})\big((\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2)\big)\boldsymbol{x}_i'\big]'\\
&\equiv 0
\end{aligned}
$$

That is

$$A + \boldsymbol{\alpha}'[B + B'] = 0$$

where

- $A = \sum_i (\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2)'(\boldsymbol{D}_1^{-1} + \boldsymbol{D}_2^{-1})(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)\boldsymbol{x}_i'$
- $B = \sum_i \boldsymbol{x}_i\big((\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2)\big)'(\boldsymbol{D}_1^{-1} + \boldsymbol{D}_2^{-1})\big((\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2)\big)\boldsymbol{x}_i'$

Therefore,

$$\boldsymbol{\alpha} = -(B + B')^{-1}A' = -2B^{-1}A'$$

which is the $\boldsymbol{\alpha}$ that maximizes the equation (4).