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

Thaddeus Tarpey[1], Lanqiu Yao [1], Eva Petkova[1] and R. Todd Ogden[2]

[1] Division of Biostatistics Department of Population Health NYU

[2] Columbia University, NY, USA

1. Problem Setting:

- In a randomized clinical trial comparing treatments to placebo for mental illnesses, there often have subjects with different groups that have similar outcomes.
- A convexity-based clustering method was developed Tarpey et al to identify sets of outcomes that are only observed in treated subjects.
- This study is aimed to improve this method by considering subjects' baseline features.

2. Solution: Maximize Purity

In this study we focus on the scenario with 2 subpopulations (population 1 and population 2). The mixed effect model with consideration of linear combination of baseline features can be expressed as:

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

- $\boldsymbol{X}$ is the design matrix of times
- $\boldsymbol{\beta}$ is the fixed effects;
- $\boldsymbol{b} \sim N(0, \boldsymbol{D})$ is the vector of random effects
- $\boldsymbol{\Gamma}(\boldsymbol{\alpha}'\boldsymbol{x})$ is the fixed effects of baseline features, where $\boldsymbol{x}$ is the matrix of baseline covariates and α is a linear transformation
- $\boldsymbol{\epsilon}$ is the random error, which is independent of the random effect

Then the its coefficient distribution is:

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

which specifies the functions' trajectories.

The Purity is defined as the optimal rule for classification in terms of minimizing the probability of misclassification. That is, the purity of point x_i with baseline features w_i is

$$p(x_i|w_i) = \frac{(\pi_1 f_1(x_i|w_i) - \pi_2 f_2(x_i|w_i))^2}{(\pi_1 f_1(x_i|w_i) + \pi_2 f_2(x_i|w_i))^2}$$

Therefore, the purity conditioning on the baseline w_i for the data is:

$$p(x|w_i) = \int \frac{(\pi_1 f_1(x_i|w_i) - \pi_2 f_2(x_i|w_i))^2}{(\pi_1 f_1(x_i|w_i) + \pi_2 f_2(x_i|w_i))^2}(\pi_1 f_1(x_i|w_i) + \pi_2 f_2(x_i|w_i) dx_i$$

p(w_i)

$$= \int \frac{(\pi_1 f_1(x_i|w_i) - \pi_2 f_2(x_i|w_i))^2}{\pi_1 f_1(x_i|w_i) + \pi_2 f_2(x_i|w_i)} dx_i$$

- The subscripts 1 and 2 mark which population the parameters are from
- $w_i = \boldsymbol{\alpha}'\boldsymbol{x_i}$
- $f_1(x_i|w_i) \sim \text{MVN}(\boldsymbol{\beta_1} + \boldsymbol{\Gamma_1}(\boldsymbol{\alpha}'\boldsymbol{x_i})), \boldsymbol{b_1})$; $f_2(x_i|w_i) \sim \text{MVN}(\boldsymbol{\beta_2} + \boldsymbol{\Gamma_2}(\boldsymbol{\alpha}'\boldsymbol{x_i})), \boldsymbol{b_2})$
- π_1, π_2 are the prior probabilities of f_1, f_2, which can be estimated as sample proportion.

Then the purity of the whole data can be estimated as:

$$\frac{1}{n}\sum_{w_i}\sum_{x_i} p(x_i|w_i)$$

The $\boldsymbol{\alpha}$ that can maximize the purity actually minimizes the probability of misclassification.

3. Convexity-Based Clustering

A general convexity based clustering method is to find a partition that maximize the function through iteration:

$$\sum_{j=1}^{k} P(B_j)\phi(E[\boldsymbol{X}|\boldsymbol{X} \in B_j])$$

where ϕ is a convex function. The algorithm can be expressed as:

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

$$D_j = \{||\lambda - h_j|| < ||\lambda - h_i||, i \neq j\},$$

where $\lambda(x) = \frac{\pi_2 f_2(x)}{f(x)}$ 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

4. Example

- Data from a 6-week longitudinal depression study. Subjects randomly asigned to Fluoxetine group v.s. placebo group. The outcome the severity of depression assessed with the Hamilton Rating Scale for Depression (HRSD)
- The purity calculation: We considered two covariates: Age, Baseline CGI. The $\boldsymbol{\alpha}$ that can achive the max purity is $\boldsymbol{\alpha} = [0, 1]$
- A Monte Carlo sample (size of 10000) was simulated to compute the probabilities $P_1(B_j)$, $P_2(B_j)$ in the clustering algorithm iteration.
- The clustering algorithm was performed without or with the baseline features combination

Table 1: Percentage of subjects classified to each cluster

	Fluoxetine, n = 196			Placebo, n = 162		
Cluster	% Responders	% Non-Responders	Total	% Responders	% Non-Responders	Total
1	29	6	35	5	0	5
2	31	14	45	24	4	28
3	4	12	16	10	31	41
4	0	4	4	0	26	26
Overall	64	36	100	39	61	100

Table 2: Percentage of subjects classified to each cluster, with consideration of baseline features

	Fluoxetine, n = 196			Placebo, n = 162		
Cluster	% Responders	% Non-Responders	Total	% Responders	% Non-Responders	Total
1	34	6	40	7	0	7
2	29	18	47	25	8	33
3	1	9	10	6	32	38
4	0	3	3	0	22	22
Overall	64	36	100	38	62	100

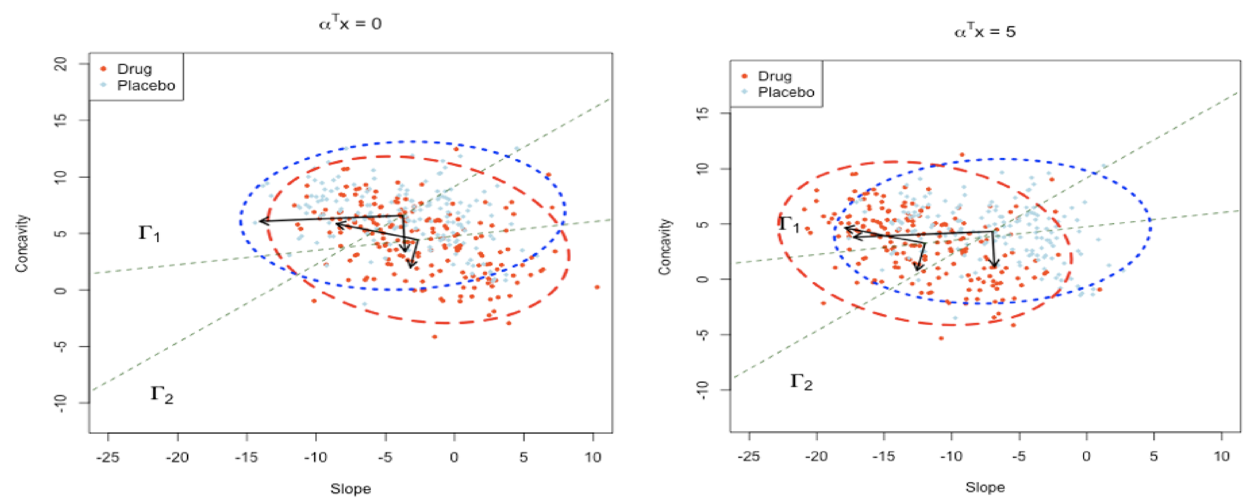

Figure 1: Contours of slope and concavity coefficients for Fluoxetine and placebo treated subjects with different $\alpha^T x$ values

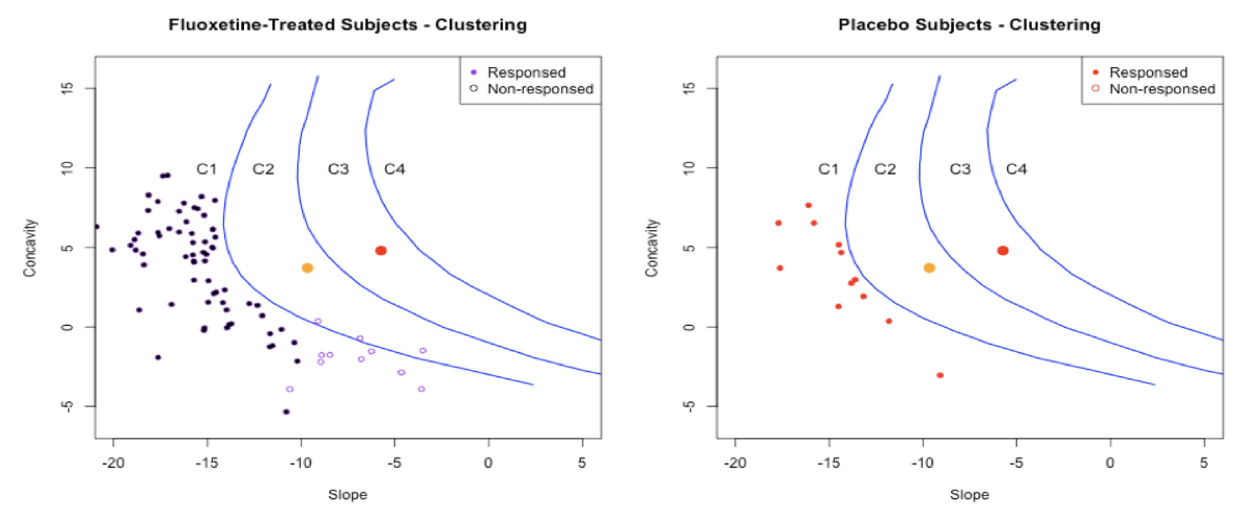

Figure 2: Convexity-based clustering partition for k = 4, subjects classified to Cluster 1

Conclusion

- To identify effective treatments when there is a lot of overlap outcomes from different groups of subjects is a challenge for mental health study.
- This study is based on the convexity-based clustering approach developed by Tarpey et al.
- By considering a linear combination of baseline biosignatures may improve the approach's performance, which can be seen from the results.
- Future work: a more efficient method for higher dimension of biosignatures is needed.

References

1. Tarpey, Thaddeus, Eva Petkova, and Liangyu Zhu, "Stratified psychiatry via convexity-based clustering with applications towards moderator analysis.," *Statistics and its interface* **9.3(2016): 255.**
2. Bock, Hans-Hermann. "Convexity-based clustering criteria: theory, algorithms, and applications in statistics." *Statistical Methods and Applications* **12.3 (2004): 293-317.**
3. Petkova, E., Tarpey, T., R. T. Ogden and Z. Su, (2014), "Generated Effect Modifiers (GEMs) in Randomized Clinical Trials," preprint.
4. Tarpey, T., R. Ogden, R. T. and Petkova, E., (2014), " A Paradoxical Result in Estimating Regression Coefficients," *The American Statistician*, in press.

Thomas R. Ten Have Symposium on Statistics in Mental Health, 2019

This work was supported by National Institute of Mental Health (NIMH grant R01 MH099003) in the USA.