Differential PP

Thursday, October 20, 2022 3:36 PM

Purpose: compare two groups of experiments from flow cytometry, compare the groups by finding the regions where they are the most different.

Data: X is [n1*p]-dim results for treatment, Y is [n2*p]-dim results for control group. Both have millions of rows for data.

Method 1:

1. Combine X and Y -> a combined large dataset, calculate the population proportion of observations, π = n1/(n1+n2)
2. Construct m data nuggets for the combined dataset
3. For each nugget, there are observations from two groups. Calculate the proportion of observations in each nugget, p_i = n_1i/(n_1i+n_2i) for i=1,2,…m.
4. For each nugget, do the test for H0: p_i = π , H1: p_i != π , then we have p values for each nugget.
5. Make a plot: x-axis is weights of nuggets/sqrt(weights)/log(weights), y-axis is -log(p-values), check the p-values' change as weights change -> purpose? Check: -log(p-values) is measurement of difference. What happens when if weights are big, two groups are very big, it could easily be significantly different. Eg, π = 44%, p_i = 43%. If the number of observations is very big, it would be significantly different. Decide **a linear boundary to decide** whether to refuse the test.
6. By p-values, we divide nuggets into two groups: one with no difference, one with significant difference. By compare values, we divide the second group into two. Finally we have three groups: one with no difference, one with p_i bigger than π , one with p_i smaller than π .
7. Reassignment: for those nuggets with |p_i - π | < 0.2, we put those into the no difference group. -> how to decide 0.2, maybe according to dataset? Consider 3 or 2 percentages. According to step 5. **Some formula?**
8. Consider Another group: different between p_i and 1-p_i -> 2p_i-1 or **1-2p_i**. If 1-2p_i > 0.8, it means in that nugget, most of observations are from one group. The nugget is almost pure.
9. Find the regions with difference, we may focus on the group with p_i bigger than π (treatment works better/activates more B-cells), conduct:
   (1). Weighted PCA -> loadings for each protein
   (2). PP
   For this sub-group?
   And another way: conduct weighted-Kmeans to get clusters for the combined data -> cluster information for all data, and then compare the group information and cluster information, eg, in each cluster, the proportions of each group -> find clusters with higher proportions -> target areas
   (consider Classification and regression tree: CART for data nuggets)

#imbalance problem between treatment and control?
#multiple treatments compare for this method?

Method 2: By differential PP

1. **For one treatment and one control: Find a d-dim projection that has a very large difference between two distributions**
   Step 1 : For each projection, (apply same projection matrix on both data), f1(y): projected data density for treatment group, f2(y): projected data density for control group, f: combined density, how to combine, simple way: f(y) = 1/2*(f1(y)+f2(y))
   Step 2: Calculate differential PP index:

$$\boxed{I = \int_{\mathbb{R}^d} (f_1(y)-f_2(y))^2 f(y)\,dy} = c\cdot\underline{\left[\int_{\mathbb{R}^d} (f_1(y)-f(y))^2 f(y)\,dy + \int_{\mathbb{R}^d} (f_2(y)-f(y))^2 f(y)\,dy\right]}$$

$$\shortparallel$$

$$\left(\begin{aligned}&\int_{\mathbb{R}^d} (f_1(y)^2-2f(y)f_1(y)+f(y)^2+f_2(y)^2-2f_2(y)f_1(y)+f(y)^2)f(y)\,dy\\ &= \int (f_1(y)^2+f_2(y)^2)f(y)\,dy - 2\int (f_1(y)+f_2(y))f(y)^2\,dy + 2\int f(y)^3\,dy\end{aligned}\right)$$

   Step 3: maximize differential PP index to find optimal projection
   Step 4: Conduct optimal projection on combined data, check two groups distribution and any area.
   Step 5: apply varimax rotation to get protein information (loadings)/ apply clustering methods to projected data to check area and protein expressions in each clusters, other way?
2. **For multiple treatments and one control: in total k groups**
   Similar things but different index: fi(y): projected data density for each group, i= 1,2,…,k, f: combined density, how to combine, simple way: f(y) = 1/k*(f1(y)+…+fk(y))

$$\int_{\mathbb{R}^d} (f_1-f_2)^2 f\,dy + \int_{\mathbb{R}^d} (f_1-f_3)^2 f\,dy + \int (f_2-f_3)^2 f\,dy = c\cdot\left[\int (f_1-f)^2 f\,dy + \int (f_2-f)^2 f\,dy + \int (f_3-f)^2 f\,dy\right]$$

$$\int_{\mathbb{R}^d} (f_1-f_2)^2 f\,dy + \int_{\mathbb{R}^d} (f_1-f_3)^2 f\,dy + \int (f_2-f_3)^2 f\,dy = c \cdot \left[\int (f_1-f)^2 f\,dy + \int (f_2-f)^2 f\,dy + \int (f_3-f)^2 f\,dy\right]$$

2. For k groups.

$$\underbrace{\sum_{\substack{i<j \\ i,j=1,\dots,k}} \int_{\mathbb{R}^d} |f_i(y)-f_j(y)|^2 f(y)\,dy}_{C_k^2 = \frac{k(k-1)}{2} \text{ integrals}} = c \cdot \underbrace{\sum_{i=1}^{k} \int |f_i(y)-f(y)|^2 f(y)\,dy}_{k \text{ integrals}}$$

Verification of equations and it could help: we do not need make pairwise comparisons between k groups (C_k^{2} = k*(k-1)/2 pairs), **just compare each group with the combined density.**

Task: find dataset to simulate and test (Davit's data or other flow cytometry data on that website), and think about things with ?