StatsM254 Statistical Methods in Computational Biology Lecture 3 - 04/08/2014

Multiple Testing Issues & K-Means Clustering

Lecturer: Jingyi Jessica Li *Scribe: Arturo Ramirez*

Multiple Testing Issues

When trying to find differentially expressed genes, we're performing a test on each gene. If we have genes $1, 2, ...m$, we're perfoming m tests.

Table 1: Decision matrix

Decisions	Accept H_0	Reject H_0	Total
True H_0	U	V	m_0
False H_0	T	S	$m - m_0$
Total	$m - R$	R	m

R is the total number of differentially expressed genes we called through the m tests. U is the number of true negatives. T is the number of false negatives (type II error). V is the number of false positives (type I error). S is the number of true positives.

Definitions related to the significance level (or type I error) of multiple tests

1: PCER

Per Comparison Error Rate (PCER) $= \frac{\mathrm{E}[V]}{m}$

2: FWER

Family Wise Error Rate (FWER) $= P(V \geq 1)$

3: FDR

False Discovery Rate (FDR)

$$\text{FDR} = \begin{cases} \mathrm{E}[\frac{V}{R}] \\ 0, \qquad \text{if } R = V = 0 \end{cases}$$

PCER

Suppose we test each of the m hypotheses at significance level α, then

$$\Rightarrow P(\text{reject } H_0^{(i)} | H_0^{(i)} \text{ is true}) = \alpha, i = 1, 2...m$$

$$\text{PCER } = \frac{\mathrm{E}[V]}{m} = \frac{\sum_{i=1}^{m} P(\text{reject } H_0^{(i)} \text{ and } H_0^{(i)} \text{ is true})}{m}$$

$$= \frac{\sum_{i=1}^{m} P(\text{reject } H_0^{(i)} | H_0^{(i)} \text{ is true}) P(H_0^{(i)} \text{ is true})}{m} \tag{1}$$

We know $P(H_0^{(i)} \text{ is true}) \in [0, 1]$, so

$$(1) \leq \frac{\sum_{i=1}^{m} P(\text{reject } H_0^{(i)} | H_0^{(i)} \text{ is true})}{m} = \frac{m \times \alpha}{m} = \alpha$$

Thus, PCER is bounded by α

FWER

$$FWER = P(V \geq 1) = P(\text{Reject at least one } H_0) = P(\bigcup_{i=1}^{m} \{\text{reject } H_0^{(i)}\}) \leq \sum_{i=1}^{m} P(\text{reject } H_0^{(i)}) \leq m \times \alpha$$

$m \times \alpha$ is a bad upper bound because it can take values that don't make sense (e.g. $m = 100, \alpha = 0.05$), but it can still give us useful information.

$$\Rightarrow \text{ If we reduce the significance level of each test to } \frac{\alpha}{m} \text{ the FWER } \leq \alpha$$

This is called the **Bonferroni Correction**

e.g. Suppose you have 1,000 genes and you want the FWER to be ≤ 0.05. Then you call the gene differentially expressed only if its p-value $\leq \frac{0.05}{1000}$

This correction is very stringent. You could miss true differentially expressed genes.

$\Rightarrow$ May result in the discovery of too few genes. What do we do?

Comment : If all $H_0^{(i)}$ are true and have level α, V $\sim Bin(m, \alpha)$

FDR (Benjamini and Hochberg, 1995)

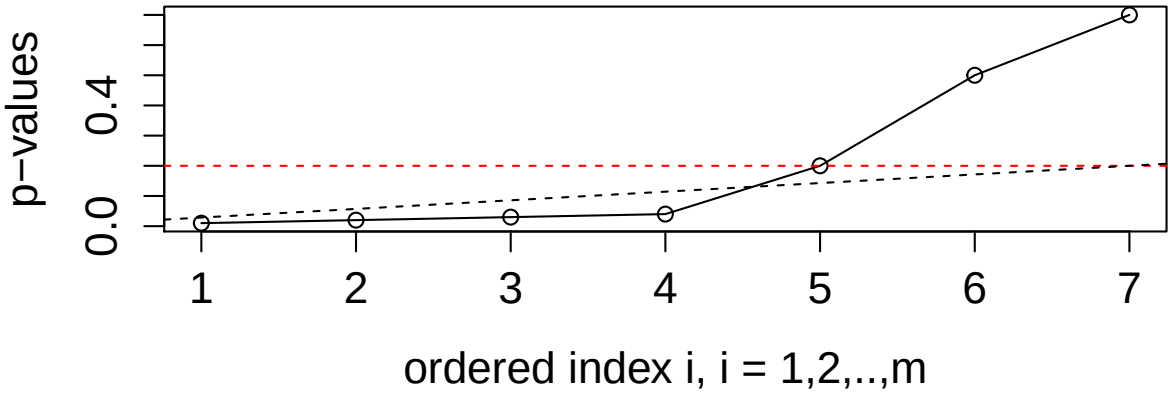
goal: control FDR = $\mathrm{E}[\frac{V}{R}] \leq \alpha$

procedure: consider testing $H_0^{(i)}, ..., H_0^{(m)}$ based on p-values $p_1, ..., p_m$.

Let's order the p-values as $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(m)}$ and their corresponding H_0's as $[H_{(1)}, H_{(2)}, ..., H_{(m)}]$

Let k be the largest i such that $p_{(i)} \leq \frac{i}{m} \times \alpha$.

Then we reject all $H_{(i)}$ for which $i \leq k$



In this example $\alpha = 0.2$ and $m = 7$. Here the red line is at $h = \alpha$ and the black line increases at $b = \frac{\alpha}{m}$

Intuition: Reject all p-values below the black line

Here you compare the i^{th} p-value to $\frac{i}{m} \times \alpha$ to make a decision $\Rightarrow$ loose

In Bonferroni you compare every p-value to $\frac{\alpha}{m}$ to make a decision $\Rightarrow$ stringent

Why does this procedure work?

If all H_0 are true, then $\mathrm{E}[V] = m \times p_{(k)} \Leftarrow$ actual significance level of each test

$$p_i = i^{th} \text{ p-value } = P(\text{reject } H_0^{(i)} | H_0^{(i)} \text{ is true})$$

Here $R = k$(we decide) $\Rightarrow \mathrm{E}[\frac{V}{R}] = \frac{m \times p_{(k)}}{k} \leq \frac{m}{k} \frac{k}{m} \times \alpha = \alpha$

Example: Account for correlation between genes

Table 2: Expression Value matrix

Gene	Condition 1	Condition 2	Test Statistic
	1-c_1 values	1-c_2 values	$\lvert S_{(1)} \rvert$
	...	...	...
	1-c_1 values	1-c_2 values	$\lvert S_{(m)} \rvert$

We've got expression values for each gene where condition 1 has c_1 replicates and condition 2 has c_2 replicates. These $c_1 + c_2$ expression values are used to calculate a test statistic $|S_i|$ for each gene (e.g. a t-statistic - the larger the value, the more likely gene i is differentially expressed). These $|S_i|$'s are then ordered from greatest to smallest in the data matrix.

$\Rightarrow$ We then permute the expression values for each gene across the conditions (i.e., permute the columns of the Expression Value Matrix) for a total of N times, each resulting in a permuted data matrix.

We want to define a c such that the genes with $|S_i|$ bellow c is called "non differentially expressed" and the genes with $|S_i|$ above c is called "differentially expressed".

Using a decision rule $|S_i| > c$ we find a total of R genes to be differentially expressed.

Question: What is the FDR?

In permutation j, define V_j as the number of genes with $|S_i|$ statistic $> c$

Estimate FDR as

$$\Rightarrow \frac{\frac{1}{N}\sum_{i=1}^{N} V_j}{R}$$

Clustering Algorithms

1: K-Means

n genes, each with an expression vector $\in \mathbb{R}^p$ (p samples)

$X_1, X_2, \ldots, X_n \in \mathbb{R}^p \Rightarrow$ assign them in to k clusters. The class label of X_i is $C(i)$ and m_k is the cluster center for the k^{th} cluster.

Objective Function (goal):

We want to minimize the total within cluster distance

$$\sum_{k=1}^{K} \sum_{C(i)=k} ||X_i - m_k||^2$$

$$\Rightarrow (\{m_k^*\}_{k=1}^K, \{C(i)^*\}_{i=1}^n)$$

$$\left(\{m_k^*\}_{k=1}^K, \{C(i)^*\}_{i=1}^n\right) = \underset{\{m_k\}_{k=1}^K, \{C(i)\}_{i=1}^n}{\operatorname{argmin}} \sum_{k=1}^{K} \sum_{C(i)=k} ||X_i - m_k||^2$$

Algorithm:

1)

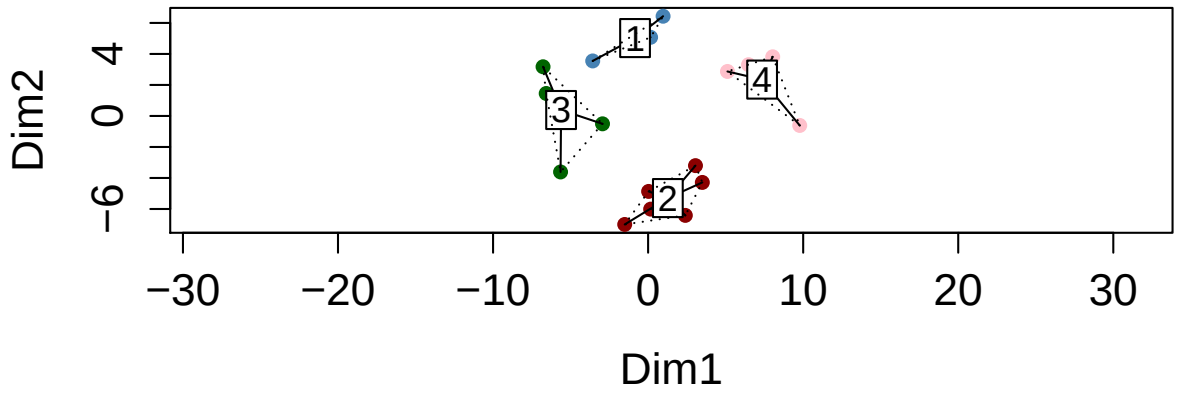
For a given cluster assignment C, minmize the total within cluster distance with respect to $\{m_k\}_{k=1}^K$

$$m_k^* = \underset{m_k}{\operatorname{argmin}} \sum_{C(i)=k} ||X_i - m_k||^2$$

$$\text{If } ||X_i - m_k||^2 = (X_i - m_k)^2 \Rightarrow m_k^* = \frac{1}{n_k} \sum_{C(i)=k} X_i$$

2)

Given the cluster centers $\{m_k\}_{k=1}^K$, minimize the total within cluster distance with regard to the cluster assignment.



e.g. 2-Dimensions. Given cluster centers 1,2,3,4, these cluster assignments would minimize within cluster distance for each cluster.

$$C(i)^* = \underset{1 \le k \le K}{\operatorname{argmin}} ||X_i - m_k||^2$$

Procedure

1) Find the optimal cluster centers given a cluster assignment; 2) Find the optimal cluster assignment given the cluster centers; 3) Iterate 1) and 2) until C converges.

Question:

Can we find the global $\min_{\{m_k\}_{k=1}^K, \{C(i)\}_{i=1}^n} \sum_{k=1}^K \sum_{C(i)=k} ||X_i - m_k||^2$ of the total within cluster distance?

This clustering algorithm is sensitive to initial cluster assignment. Not every initial cluster assignment can lead to the global min.

Solution:

Use multiple initial cluster assignments to check if there is a common solution reached by a majority of the initial assignment schemes.