STAT 530 Bioinformatics: Homework 3

Due Mar 13, 2017

For problems using R, turn in your answers in the form of a compiled R notebook PDF.

Problem 1 (5 points)

Suppose we want to perform FDR adjustment on a very large number of p-values. Suppose we have m total p-values $P_1, \ldots, P_m$. The jth FDR-adjusted p-value is defined as

$$P_j^\star = \min_{P_{j'} \geq P_j} \frac{m P_{j'}}{\sum_k I(P_k \leq P_{j'})}.$$

The minimim is taken over all observed p-values that are larger than P_j.

However, in some cases, such as in eQTL analysis, there are so many p-values that it is too cumbersome to store all of them, and instead only the smallest ones are kept. Suppose we have only stored the p-values less than or equal to some threshold τ. In this case we typically take the FDR-adjusted p-value to be

$$P_j^\dagger = \min_{\tau \geq P_{j'} \geq P_j} \frac{m P_{j'}}{\sum_k I(P_k \leq P_{j'})}.$$

This is what the R function `p.adjust` does. Suppose you conducted m total tests but only stored the smallest p-values in a vector `x`. Then `p.adjust(x,method=''fdr'',n=m)` will return FDR-adjusted p-values according to the formula above.

Fix a p-value P_j that is less than τ and is small enough to be included in `x`. For this j, will $P_j^\star$ be equal to $P_j^\dagger$? If not, what is the relationship between $P_j^\star$ and $P_j^\dagger$? Hint: test this out by simulating some examples in R.

Problem 2 (5 points)

In the slides we showed that in some cases, adjusting for more tests using FDR can give more significant discoveries. To be specific, suppose we have two sets of p-values, A and B. Suppose we apply FDR only to p-values in A, and we discover D significant p-values from A at FDR level α. We showed that in some cases, if we combine p-values from both A and B and then apply FDR at level α, we can sometimes discover more than D significant p-values from A. Under what heuristic condition on the p-values in B will this be possible?