High-Dimensional Data Analysis (BIOS 7600)
Breheny

Assignment 2
Due: Monday, February 15

1. Derive the relationship between FDR and local FDR on slide 18 of the 2-1 notes:

$$\mathbb{E}\{\phi(z)|z \in \mathcal{Z}\} = \phi(\mathcal{Z}),$$

where $\phi(z)$ is the local FDR at point z and $\phi(\mathcal{Z})$ is the FDR over the set $\mathcal{Z}$. To be clear, this problem involves the true FDR and local FDR – no estimates are involved.

2. (a) Can the Benjamini-Hochberg FDR be less than the p-value? In other words, is it possible for, say, a test with a p-value of .06 to be rejected at a false discovery rate of 5%? Either prove that this cannot happen, or provide a counterexample.
   (b) Same as (a), only what if we estimate π_0?
   (c) Same as (a), only for local false discovery rates in which we've estimated δ_0 and σ_0. To keep this problem distinct from (b), assume that $\pi_0 = 1$.

3. Show that if a random variable X satisfies $cX \sim \chi^2_\nu$, then

$$X \sim \text{Gamma}\Big(\frac{\nu}{2}, \frac{c}{2}\Big),$$

where $\text{Gamma}(\alpha, \beta)$ denotes the gamma distribution with shape parameter α and rate parameter β.

4. Simulate 1,000 z-statistics according to the following scheme:

$$\begin{aligned} Z_i &\sim \text{N}(3,1) && \text{for } i = 1, 2, \ldots, 100 \\ Z_i &\sim \text{N}(0,1) && \text{for } i = 101, 102, \ldots, 1000 \\ \text{Cor}(Z_i, Z_j) &= 0.1 && \text{if } i, j > 100 \\ \text{Cor}(Z_i, Z_j) &= 0 && \text{otherwise} \end{aligned}$$

Use the Benjamini-Hochberg procedure with the known value $\pi_0 = 0.9$ to reject as many hypotheses as possible subject to an FDR cutoff of 20%, then calculate the actual false discovery rate among those rejected hypotheses.

   (a) Repeat this procedure 2,000 times. Calculate the average true FDR and the standard deviation of the true FDRs.
   (b) Same as (a), but without any correlation between the z-statistics (i.e., $\text{Cor}(Z_i, Z_j) = 0 \, \forall \, i, j$). Again, what are the mean and standard deviation of the true FDRs across the 2,000 replications?
   (c) Produce a figure (e.g., a pair of histograms or a boxplot) comparing the true FDRs from (a) and (b), and comment on the effect of correlation upon FDR control.

5. In the Golub leukemia data, there are some rather large outliers among the raw gene expression values. Re-analyze the data using Wilcoxon rank-sum tests instead of t-tests, and obtain an estimator for π_0. Comment on whether the estimated fraction of null features appears to be sensitive to the type of test in this example.

6. The course website contains gene expression data from a case-control study of prostate cancer (`Singh2002`). Carry out a t-test for differential expression between cases and controls, and carry out the following multiple comparison adjustment procedures: Bonferroni, Holm, FDR (Benjamini-Hochberg), FDR with estimation of π_0, and local FDR. For each adjustment, calculate (a) the z-value cutoff corresponding to a 10% error rate, and (b) the number of differentially expressed genes that exceed that threshold. Report these numbers in a table. For the local false discovery rate, there are a variety of choices and software packages you could use; do whatever you feel is appropriate, but you must describe the approach you used.

7. The course website contains a data set from an experiment to identify genes in ER+ breast cancer cells that respond to estrogen (`Scholtens2004`); a more detailed description of the experimental design is available online. Analyze the data to produce three lists of genes: genes that respond to estrogen, "early responders" for which the estrogen response is stronger in the short term than it is later, and "late responders" for which the estrogen response is stronger later than it is in the first 10 hours.

   For this assignment, write up a brief "Methods" and "Results" section, as it might appear in a scientific journal, each consisting of one or possibly two paragraphs, describing what you did (Methods) and what you found (Results). For the methods section, you must use the empirical Bayes approach we discussed in class on February 3, but everything else is up to you – there are a variety of reasonable ways you could interpret the scientific questions. For the results section, describe the number of genes you found in each category, the cutoff criterion you used, and a list or small table of 2 or 3 representative genes from each category so that I can check whether your results make sense.