High Dimensional Data analysis (BIOS:7240)
Breheny

Assignment 2
Due: Monday, February 15

1. *Relationship between FDR and local fdr.* Derive the relationship between FDR and local FDR on slide 17 of the "Local false discovery rates" notes:

$$\mathbb{E}\{\text{fdr}(z)|z \in \mathcal{Z}\} = \text{Fdr}(\mathcal{Z}),$$

where $\text{fdr}(z)$ is the local FDR at point z and $\text{Fdr}(\mathcal{Z})$ 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. *Scaled χ^2.* 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.

3. *FDR estimates when $z = 0$.* For this problem, let z denote the z-statistic of a two-sided test. For (a)-(d) below, you do not have to provide a proof; simply stating the answer is fine.

   (a) In the Benjamini-Hochberg approach, what is the q-value for a feature with $z = 0$?

   (b) What is the q-value for a feature with $z = 0$ if we estimate π_0?

   (c) For the Gaussian mixture model approach, what is the range of values that fdr could have, for a feature with $z = 0$? Hint: Consider the limiting cases where $\sigma_k^2 \to 0$ and $\sigma_k^2 \to \infty$.

   (d) For the Gaussian mixture model approach, what is the range of values that fsr can have, for a feature as $z \to 0^+$?

4. *FDR accuracy in the presence of correlated tests.* 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. *Prostate cancer study.* 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 apply out the following multiple comparison adjustment procedures with an error rate (FWER/FDR/local FDR/local FSR) of 10%:

   - Bonferroni
   - Holm
   - FDR (Benjamini-Hochberg)
   - FDR with estimation of π_0
   - Local FDR (there are a variety of choices and software packages you could use; do whatever you feel is appropriate, but describe the approach you used)

   For each method, report both the number of significant results and the smallest z-statistic (in absolute value) that was still considered significant.

6. *Breast cancer gene expression in response to estrogen.* 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 moderated testing approach we discussed in class, where the variance estimates are borrowed across genes, but everything else is up to you – in particular, there are a variety of reasonable ways you could interpret the scientific questions and address the multiple testing issue. For the results section, describe the number of genes you found in each category, the 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.