STAT 30850 Project Proposal

Arjun Biddanda (abiddanda@uchicago.edu)
Joseph Marcus (jhmarcus@uchicago.edu)

February 14, 2016

Overview

In many scenarios we can imagine that data is constantly streaming in as a result of a time-dependent generating process. We would like to conduct hypothesis testing on the data points a s they are streaming in, while still retaining a adequate FDR control. Our proposal is to develop a Bayesian method for FDR control in this online setting.

Initial Proposal of Model

Suppose that time $t \in \{0, 1, ..., \tau, ...\}$, and that we have statistcs X_t independently coming from the following mixture distribution:

$$X_t \sim \pi_0 \cdot N(\mu_0, \sigma_0^2) + (1 - \pi_0) \cdot N(\mu_1, \sigma_1^2)$$

We will further simplify this and obtain the p-values for each of the data points P_t. We wish to use all datapoints prior to τ to estimate the proportion of nulls (π_0) before we switch to a Bayesian FDR method (such as Storey's method). Note that as we proceed through the data stream, all of the p-values $P_0, ..., P_t$ can be used to update our estimate of the true proportion of nulls. Note that in the above setting we assume that the true proportion of nulls (π_0) does not vary with time. We also define the rejection region of the p-values as $[0,\gamma]$.

Thus our initial estimate of the proportion of nulls will be (according to Storey):

$$\hat{\pi}_0^\tau(\lambda) = \frac{\#\{p_i > \lambda\}}{(1 - \lambda)\tau}$$

And more generally our updated estimates will be:

$$\hat{\pi}_0^t(\lambda) = \frac{\#\{p_i > \lambda\}}{(1 - \lambda)t}$$

Then we can define our similar data-adaptive FDR and pFDR estimators at time $t > \tau$ as:

$$FDR_\lambda^t(\gamma) = \frac{\hat{\pi}_0^t(\lambda)\gamma}{(1 - \lambda)(R(\gamma) \vee 1)}$$

$$pFDR_\lambda^t(\gamma) = \frac{\hat{\pi}_0^t(\lambda)\gamma}{(1 - \lambda)(R(\gamma) \vee 1)(1 - (1 - \gamma)^t)}$$

Questions Proposed

- τ : which time-step to switch to the Bayesian FDR model? What model to use prior to this model? Or should we not even switch and just reject everything before τ?
- π_0 : How many nulls are there relative to the signals within the data? Are we able to detect the signals even when there are very few of them?
- $\mu_1 >> \mu_0$: The relative strength of the signals vs. the nulls. How weakly can we classify signals?
- Calculating λ_{best} : Storey formulates a method (Section 9) to compute the optimal value of λ for a given set of data. How could recalculating this value according to streaming data affect FDR? Recalculate for most recent "chunk" of time or aggregate through time?
- Clustering of Signals : How does the method react to clusters of signals together?

References

1. Storey, John. *A direct approach to false discovery rates.* 2002. *Journal of the Royal Statistical Society*