STAT 30850 Project Proposal

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

February 18, 2016

Overview

There are many scenarios where hypothesis testing is peformed sequentially as new data becomes availialbe through time. Streaming data is common in many modern applications such as high frequency stock trading, novel results from clinical trials, and large-scale adverstisment experiments conducted by Google and Facebook. In this context one would like to test multiple hypothesis on the data as its streaming in, while still retaining adequate FDR control, similalry to approaches developed for non-streaming data. Here we propose to extend and apply Bayesian methods for FDR control in an online testing setting.

Setup

Suppose that time $t \in \{0, 1, \ldots, \tau\}$, and that test statistcs X_t, from some experiment, are independently and indentically distributed under the mixture model:

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

For instance we could imagine X_t are Z scores such that $\mu_0 = 0$ and $\sigma^2 = 1$. Our goal is to estimate the mixture model $\{\pi_0, \mu_1, \sigma_1\}$ for each t as we oberseve a new X_t. We plan to control for FDR at the appropriate level α under the estimated mixture model at time t , for simplicity call it M_t.

Approach

We propose to use Markov Chain Monte Carlo (MCMC) to sample from the posterior distributions of $\{\pi_0, \mu_1, \sigma_1\}$ allowing us to do inference of M_t and providing measures of uncertity in these parameters at each time step or at some interval of time. To this end we set priors distributions on $\{\pi_0, \mu_1, \sigma_1\}$:

$$\pi_0 \sim Beta(1, 1)$$
$$\mu_1 \sim Normal()$$
$$\sigma_2 \sim Normal()$$

If Z is an indicator for X_t being a signal then $P(Z = 0) = \pi_0$ and $P(Z = 1) = 1 - \pi_0$. From the above mixture we know:

$$X_t \mid Z = 0 \sim N(0, 1)$$

$$X_t \mid Z = 1 \sim N(\mu_1, \sigma_1)$$

If we observe t test statistics at time t we can compute the likelihood of the model as:

$$L(M_t) = P(X_1, \ldots, X_t \mid \pi_0, \mu_1, \sigma_1) = \prod_{i=1}^{t} (\pi_0 \cdot P(X_i \mid Z = 0) + (1 - \pi_0) \cdot P(X_i \mid Z = 1))$$

From bayes theorom $P(\pi_0, \mu_1, \sigma_1 \mid X_1, \ldots, X_t) \propto P(\pi_0) \cdot P(\mu_1) \cdot P(\sigma_1) \cdot P(X_1, \ldots, X_t \mid \pi_0, \mu_1, \sigma_1)$

Initial Proposal of Model

Suppose that time $t \in \{0, 1, ..., \tau\}$, and that test statistcs X_t are independently distributed under the mixture model:

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

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,γ].

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*

- interval for doing mcmc (gibbs?)
- think about how
- how frequently you want to restimate
- doing mcmc on one