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 conduncted 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 $t \in \{0, 1, \ldots, n\}$, 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 $\sigma_0^2 = 1$. Our goal is to estimate the mixture model $\{\pi_0, \mu_1, \sigma_1^2\}$ for each t as we observe 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 $\{\pi_0, \mu_1, \sigma_1^2\}$ 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 $\{\pi_0, \mu_1, \sigma_1^2\}$:

$$\pi_0 \sim Beta(\alpha, \beta)$$

$$\mu_1 \sim Normal(\mu_s, \tau_s)$$

$$\sigma_1^2 \sim Inverse - Gamma(\alpha^*, \beta^*)$$

If Z is a latent 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:

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

$$X_t \mid Z = 1 \sim N(\mu_1, \sigma_1^2)$$

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

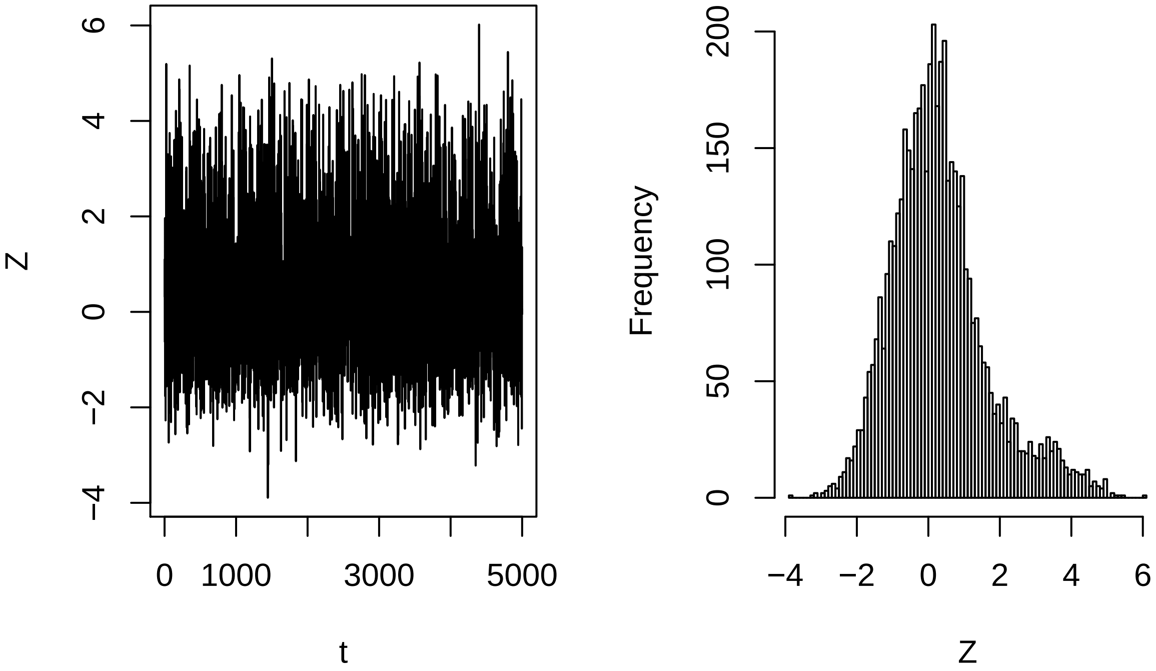
$$L(M_t) = P(X_1, \dots, X_t \mid \pi_0, \mu_1, \sigma_1) = \prod_{i=1}^{t} (\pi_0 \cdot P(X_i \mid Z = 0) + (1 - \pi_0) \cdot P(X_i \mid Z = 1))$$

From Bayes Theorem

$$P(\pi_0, \mu_1, \sigma_1 \mid X_1, \dots, X_t) \propto P(\pi_0) \cdot P(\mu_1) \cdot P(\sigma_1) \cdot P(X_1, \dots, X_t \mid \pi_0, \mu_1, \sigma_1)$$

We plan to sample from this posterior distribution using a component-wise Metropolis Hasting algorithim with a symetric proposal distribution and acceptence ratios defined from the above priors and likelihood. We plan to test approach emprically via simulations where we know the true mixture component distributions and propritions as seen below.

Z-Scores Simulated from Mixture



Above is a time-series of independent samples of Z-scores that are from the mixture distribution with parameters $\pi_0 = 0.90, \mu_1 = 3, \sigma_1^2 = 1$.

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? (i.e. what if there is some time-dependency in the transitions possible?)

References

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