

STAT 30850 Project Proposal

Arjun Biddanda (abiddanda@uchicago.edu)

Joseph Marcus (jhmarcus@uchicago.edu)

February 29, 2016

Overview

There are many scenarios where hypothesis testing is performed sequentially as new data becomes available through time. Streaming data is common in many modern applications such as high frequency stock trading, novel results from clinical trials, and large-scale advertisement experiments conducted by Google and Facebook. In this context one would like to test multiple hypotheses on the data as it is streaming in, while still retaining adequate FDR control, similarly 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 $t \in \{0, 1, \dots, n\}$, and that test statistics X_t , from some experiment, are independently and identically 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 uncertainty in these parameters at each time step or at some interval of time. To this end we set priors distributions on $\{\pi_0, \mu_1, \sigma_1^2\}$:

$$\pi_0 \sim \text{Beta}(\alpha, \beta)$$

$$\mu_1 \sim \text{Normal}(\mu_s, \tau_s)$$

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

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:

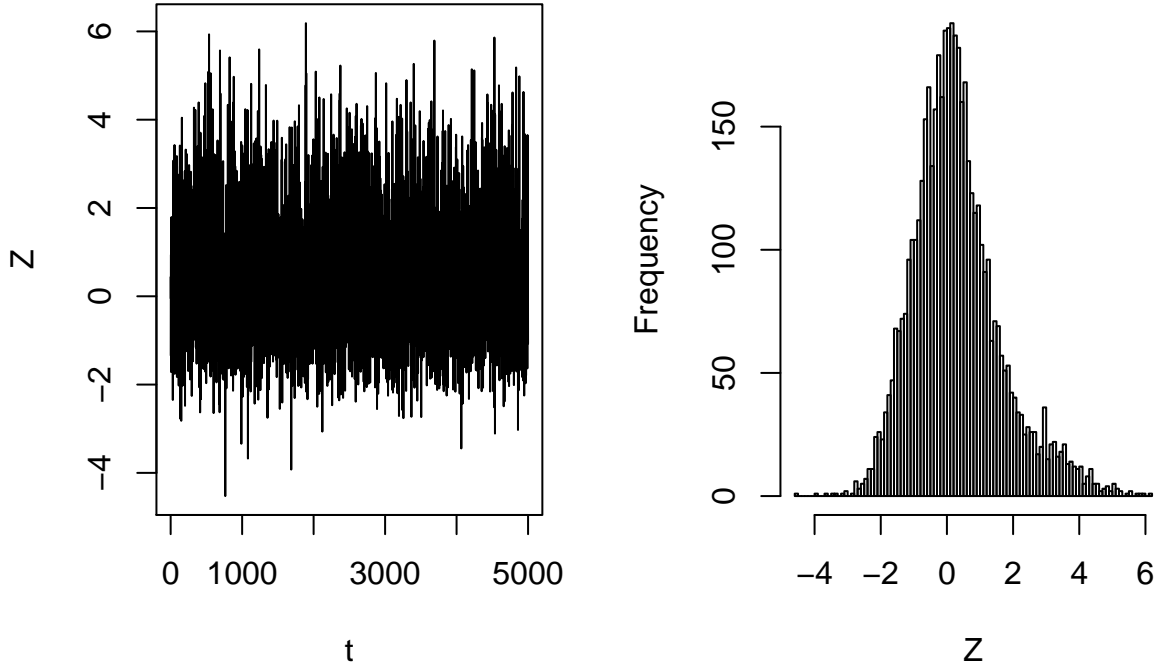
$$L(M_t) = P(X_1, \dots, X_t \mid \pi_0, \mu_1, \sigma_1) = \prod_{i=1}^t (\pi_0 \cdot P(X_i \mid Z = 0) + (1 - \pi_0) \cdot P(X_i \mid Z = 1))$$

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 algorithm with a symmetric proposal distribution and acceptance ratios defined from the above priors and likelihood. We plan to test this approach empirically via simulations where we know the true mixture component distributions and proportions 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

- Exploring efficient ways to sample from posterior of the mixture. Potential to sample periodically from the time series, rather than every time-step.
- Setting α^* to have low FDR at beginning ($t < \tau$), because of uncertainty in inferred mixture model. Explore methods to be conservative based on credible intervals of the posterior distributions of the parameters.
- 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?
- $\mu_1 \gg 0$: The relative strength of the signals vs. the nulls. How can we classify weak signals?
- Comparison of our Bayesian method against LBOND/LBORD or a sequential version of Storey's method.
- Clustering of Signals : How does the method react to clusters of signals together? Varying levels/direction of correlation amongst signals.