STAT 30850 Final Report

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

March 17, 2016

Introduction

There are many contexts where data is observed as a incoming stream of datapoints sequentially through time. In high frequency stock trading, investment firms have to make rapid decisions in response to new stock evaluations within hundredths of a second. Another example is when technology companies test the effects of advertisements on the "click behavior" of a user through A/B testing, which measures the effectiveness of the advertisement (Kohavi et al. 2009, Javanmard and Montanari (2015), Aharoni and Rosset (2014)). The setting in which hypothesis testing must be performed on streaming data, with only access to the previous data, is called "online testing".

In online testing, controlling the False Discovery Rate (FDR) at a given level has unique challenges as one does not observe all the data beyond the current time point. Here we propose a Bayesian approach to control FDR in the online testing setting. We review and contrast our method to previously developed algorithms that control FDR in online hypothesis testing, but are generally conservative (low power). We show our approach has higher power when compared to previous methods, while still maintaining FDR control through empirical simulations. Finally, we discuss future extensions, caveats, and applications of our method.

Background

Broadly speaking, previous methods for controlling FDR in the online testing context use heuristics that increase or decrease the level at which one rejects a test depending on the number of previous discoveries made. Here we review three related approaches to FDR control: α-investing, Levels Based on the Number of Discoveries (LBOND), Levels Based on Recent Discoveries (LBORD) (Foster and Stine 2007, Javanmard and Montanari (2015)).

α-investing

Let:

t - be a time index
$w(t)$ - be a wealth function which changes through time
P_t - be a p-value output from an arbitrary test at time t
α - a global level for FDR control
α_t - a time specific level for FDR control

In α-investing, one defines a wealth function w representing the amount of "wealth" or how much we can "afford" a false-positive. We imagine p-values are streaming to the researcher/statistician over time t which are provided by some arbitrary test. We then proceed to run the α-investing procedure:

1. Set $w(t=0) = \alpha$
2. At time t choose $\alpha_t \leq \frac{w(t-1)}{1+w(t-1)}$
3. Reject the null hypothesis if $P_t \leq \alpha_t$
4. Define $w(t)$ as a function of $w(t-1)$

$$w(t) = \begin{cases} w(t-1) + \alpha & P_t \leq \alpha_t \\ w(t-1) - \frac{\alpha_t}{1-\alpha_t} & P_t > \alpha_t \end{cases}$$

5. Repeat the procedure starting back at (2) for time $t+1$.

As we can see above when we reject the null, the wealth function grows and when we fail to reject the null the wealth function decays. Specifically at time 0 we set the wealth function to a "global level" α. We then proceed to set a time specific α_t. We then reject or fail to reject the p-value P_t from time t and redefine our wealth function w depending on what decision was made.

This ensures that the more discoveries we make the less stringent we are (gaining wealth) and reciprocally the fewer discoveries we make we will be more stringent when deciding to reject a future datapoint (losing wealth). For instance, if we fail to reject the null hypothesis for a long stretch of time, a p-value must be exceptionally low to overcome the current state of the wealth function.

LBOND / LBORD

Let:

t - be a time index
P_t - be a p-value output from an arbitrary test a time t
α - a global level that one would like to control FDR at
β_t - a time specific weight
D_t - count of discoveries made up to time t

In LBOND we define a series of weights β_t which sum up to the global level α. We then set a time specific α_t equal to the weight at time t multiplied by the max of 1 and the number of discoveries made up to the last time step D_{t-1}. We reject a p-value P_t if it is less than α_t and add to our discovery count.

1. At time t set $\alpha_t = \beta_t \cdot max\{1, D_{t-1}\}$ where $\sum_{t=1}^{\infty} \beta_t = \alpha$
2. Reject if $P_t \leq \alpha_t$

3. If $P_t \leq \alpha_t$ add to D
4. Repeat procedure starting at (1) for time $t+1$

Levels Based on Recent Discoveries follows a similar approach but uses weights from the time when the last discovery was made.

1. At time t set $\alpha_t = \beta_t \cdot max\{1, D_{t-\tau(t)}\}$ where $\sum_{t=1}^{\infty} \beta_t = \alpha$
2. Reject if $P_t \leq \alpha_t$
3. If discovery add to D
4. Repeat procedure starting at (1) for time $t+1$

Where $\tau(t)$ is the time of the most recent discovery before time t and $\tau(t=1) = 0$. LBORD has consistent power over time because the β weight is reset after each discovery.

Methods

Bayesian FDR

Here we propose to apply a Bayesian approach to FDR control in the online testing setting. Specifically we follow the work of Efron and model our streaming data as test statistics coming from a mixture model (Efron and Tibshirani 2002). A Bayesian approach to FDR control considers an underlying mixture distribution consisting of a *null* and *signal* components. The null component represents the distribution of the test statistic under the null hypothesis, and the signal component represents the distribution of the statistic under the alternate hypothesis.

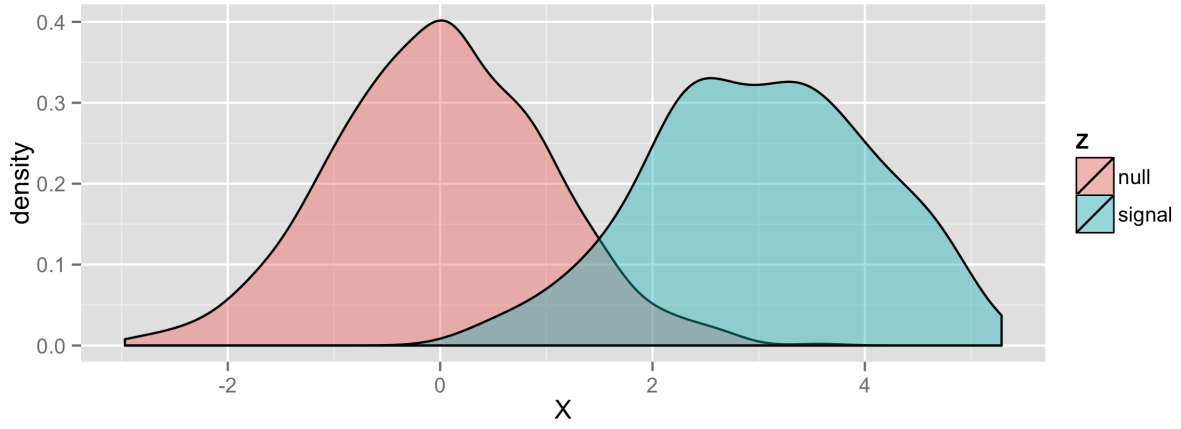

Figure 1: Density of a mixture distribution of Gaussians simulated with 80% proportion of nulls, the mean of the signal component at 3, variance of the signal component at 1, the mean of the null component at 0 and the variance of the null component at 1

Let:

X - be a test statistic
π_0 - be the proportion of nulls
μ_1 - be the mean of the signal component
σ_1^2 - be the variance of the signal component

X can be modeled as a mixture of Gaussians:

$$X \mid \pi_0, \mu_1, \sigma_1^2 \sim \pi_0 N(0,1) + (1-\pi_0) N(\mu_1, \sigma_1^2)$$

In *Figure 1* we can see a plot the resulting density of a simulated mixture model with the underlying parameters $\theta = \{\pi_0 = 0.8, \mu_1 = 3, \sigma_1^2 = 1\}$. We can see that assuming the data comes from an underlying mixture model with diverged means between the signal and null components can inform flexible approaches to control FDR. Particularly, we assume that we only know the parameters of the null component. We can find the Bayesian interpretation of FDR:

$$\begin{aligned}
\widehat{FDR}(x) &= \mathbb{E}[\widehat{FDP}(x)] \\
&= \mathbb{E}\left[\frac{P(X \in H_0, X > x)}{P(X > x)}\right] \\
&= \mathbb{E}\left[\frac{\pi_0(1-\Phi(x))}{(\pi_0(1-\Phi(x)) + (1-\pi_0)(1-\Phi(\frac{x-\mu_1}{\sigma_1}))}\right] \\
&= \frac{\pi_0(1-\Phi(x))}{(\pi_0(1-\Phi(x)) + (1-\pi_0)(1-\Phi(\frac{x-\mu_1}{\sigma_1}))}
\end{aligned}$$

If we know π_0, μ_1, and σ_1^2, then we can control FDR at a given level:

$$\alpha = \frac{\pi_0(1-\Phi(\hat{x}))}{\pi_0(1-\Phi(\hat{x})) + (1-\pi_0)\left(1-\Phi\left(\frac{\hat{x}-\mu_1}{\sigma_1}\right)\right)}$$

Where we reject X if $X > \hat{x}$.

Markov Chain Monte Carlo (Gibbs Sampler)

We apply this mixture model framework to online testing by estimating the unknown parameters of the Gaussian mixture model described above at each time point t. Specifically, we use a Markov Chain Monte Carlo approach to sample from the posterior distributions of the unknown parameters $\theta = \{\pi_0, \mu_1, \sigma_1^2\}$.

Let:

t - time index of a test statistic streaming in
X - a vector of t test statistics that have streamed in
X_t - the test statistic at the t^{th} time point
Z - vector of latent states of X_t being a signal or null
Z_t - latent state at time t of X_t being a signal or null
π_0 - proportion of nulls
μ_1 - mean of the signals
σ_1^2 - variance of the signals

As described above we model X_t as a mixture of Gaussians:

$$X_t \mid \pi_0, \mu_1, \sigma_1^2 \sim \pi_0 N(0,1) + (1-\pi_0)N(\mu_1, \sigma_1^2)$$
$$X_t \mid Z_t = 0 \sim N(0,1)$$
$$X_t \mid Z_t = 1, \mu_1, \sigma_1^2 \sim N(\mu_1, \sigma_1^2)$$

We can reparameterize this model in terms of the precision ϕ_1 of the signals and write down the likelihood conditioned on the latent indicators as:

$$L(\pi_0, \mu_1, \sigma_1^2 \mid X, Z) \propto (\pi_0)^{n_0} exp(-\frac{1}{2}\sum_{t:z_t=0} x_t^2) \cdot (1-\pi_0)^{n_1} exp\left(-\frac{\phi_1}{2}\sum_{t:z_t=1}(x_t-\mu_1)^2\right)$$

where n_0 and n_1 are the number of observed nulls and signals respectively. We can then set priors on π_0, μ_1, ϕ_1 which satisfy conjugacy:

$$\pi_0 \sim Beta(\alpha, \beta)$$
$$\phi_1 \sim Gamma(\frac{a}{2}, \frac{b}{2})$$
$$\mu_1 \mid \phi_1 \sim Normal(\mu^*, \frac{1}{\alpha^*\phi_1})$$

thus the posterior distributions of these parameters can be written as:

$$\pi_0 \mid X, Z = 0 \sim Beta(\alpha + n_0, \beta + n_1)$$
$$\phi_1 \mid X, Z \sim Gamma\left(\frac{a+n_1}{2}, b + \sum_{t:z_t=1}(x_t-\mu_1)^2\right)$$
$$\mu_1 \mid X, Z, \phi_1 \sim Normal\left(\frac{\alpha^*\mu^* + n_1 + \bar{x}_1}{\alpha^* + n_1}, \frac{1}{(\alpha^*+n_1)\phi_1}\right)$$

We also need to sample from the posterior of Z due to the conditional dependencies above:

$$P(Z_t \mid X_t = x_t, \pi_0, \mu_1, \phi_1) = \frac{\pi_0 exp(-\frac{x_t^2}{2})}{\pi_0 exp(-\frac{x_t^2}{2}) + ((1-\pi_0)\phi_1 exp(-\frac{\phi_1}{2}(x_t - \mu_1)^2))}$$

We proceed to run the Gibbs sampling algorithm as follows:

1. Set $\pi_0^{(0)}$, $\mu_1^{(0)}$ and $\phi_1^{(0)}$
2. Update Z by sampling from its posterior conditioned on X and the current values π_0, μ_1, and ϕ_1
3. Update π_0 by sampling from its posterior conditioned on X
4. Update ϕ_1 by sampling from its posterior conditioned on X and the current value of Z
5. Update μ_1 by sampling from its posterior conditioned on X and ϕ_1
6. Repeat from (2) for n iterations

Results

Simulation Study

We simulated independent and identically distributed data from a mixture model with the following parameters $\theta = \{\pi_0 = 0.8, \mu_1 = 3, \sigma_1^2 = 1\}$ for 1000 time-steps. The resulting simulated values can be seen in *Figure 2*.

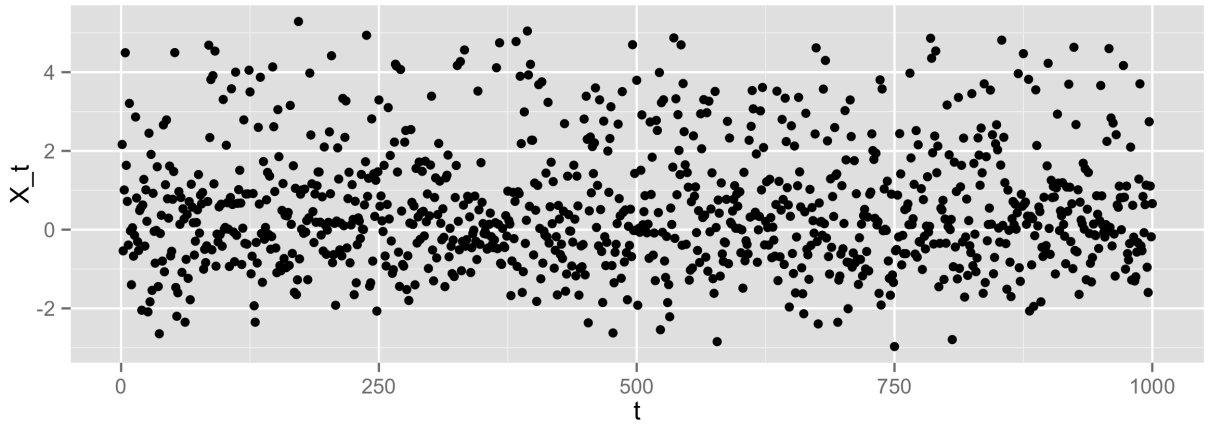

Figure 2: 1000 Simulated Z-scores from a mixture distribution with $\theta = \{\pi_0 = 0.80, \mu_1 = 3, \sigma_1^2 = 1\}$

From these simulated values we applied our Gibbs sampler to perform parameter estimations at each time point. The algorithm proceeds as follows:

1. At timestep t, use $X = (X_1, ..., X_t)$
2. Run 1000 iterations of the Gibbs Sampler using X, inculding the current datapoint as well as all of the ones from the past.
3. After a burn-in period of 50 iterations, we then choose one sample every 10 iterations.
4. Calculate the 95% credible interval from the posterior samples of the parameters
5. Repeat for the $t + 1$ timestep

In *Figure 3* we can see the 95% credible intervals of the parameters as a function of the time. Note that we *a priori* would expect the parameter estimation in the beginning to be poor due to a small amount of data used for the Gibbs sampling to estimate the posterior distribution.

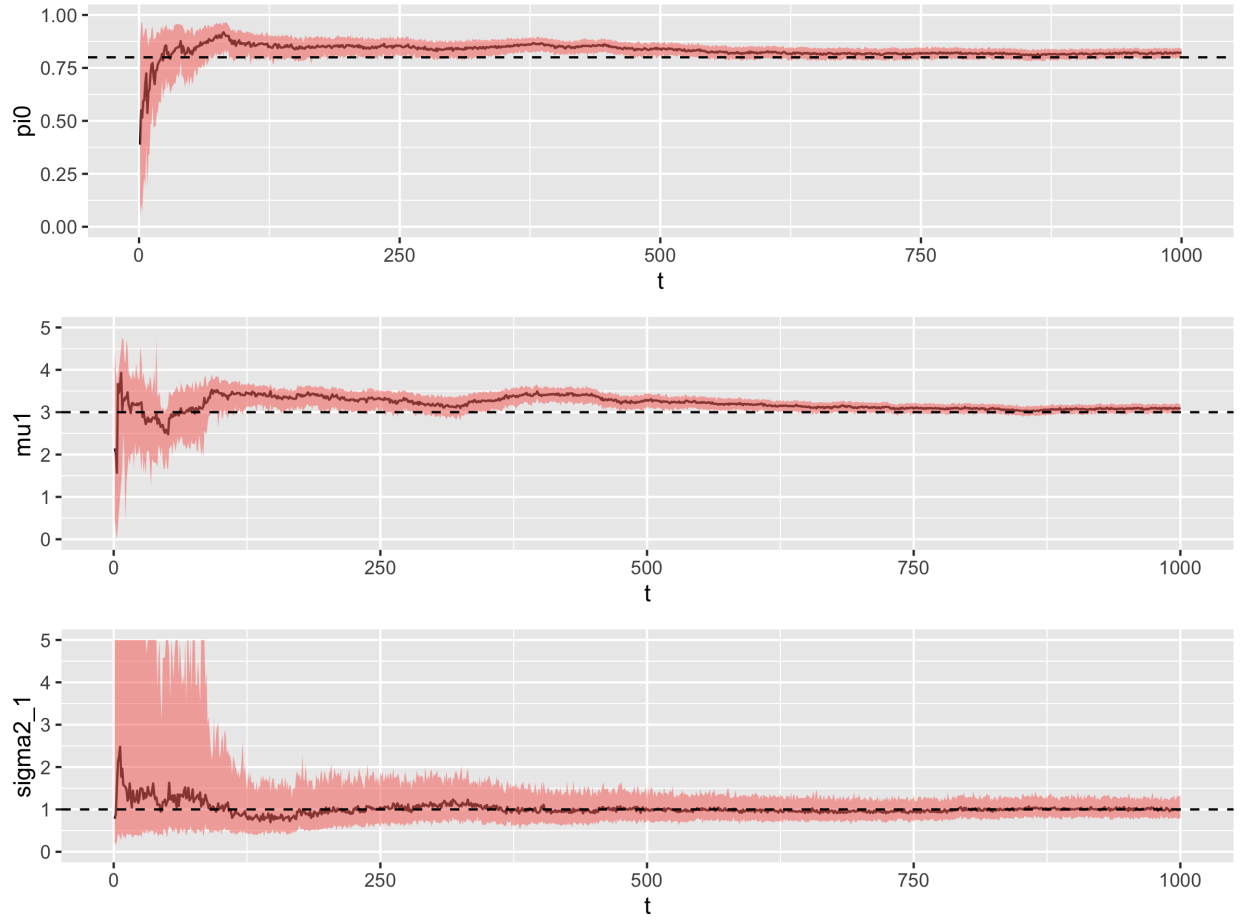

Figure 3: Sequentially estimated credible intervals of the mixture model parameters under simulation conditions. Note that for σ_1^2 we truncated the 95^{th} quantile to 5 due to numerical precision.

Estimating $\hat{x}$ Conservatively

Intuitively in the beginning of the time-series we have very little data to estimate our model, and thus our Bayesian FDR threshold may not be valid if we take a summary of our parameter estimates such as the posterior mean. Thus we performed some numerical experiments in order to determine the appropriate bounds of $\theta = \{\pi_0, \mu_1, \sigma_1^2\}$. We simply varied one of the parameters while keeping the others fixed to the parameters in our simulation. The results of these simulations are shown in *Figure 4*

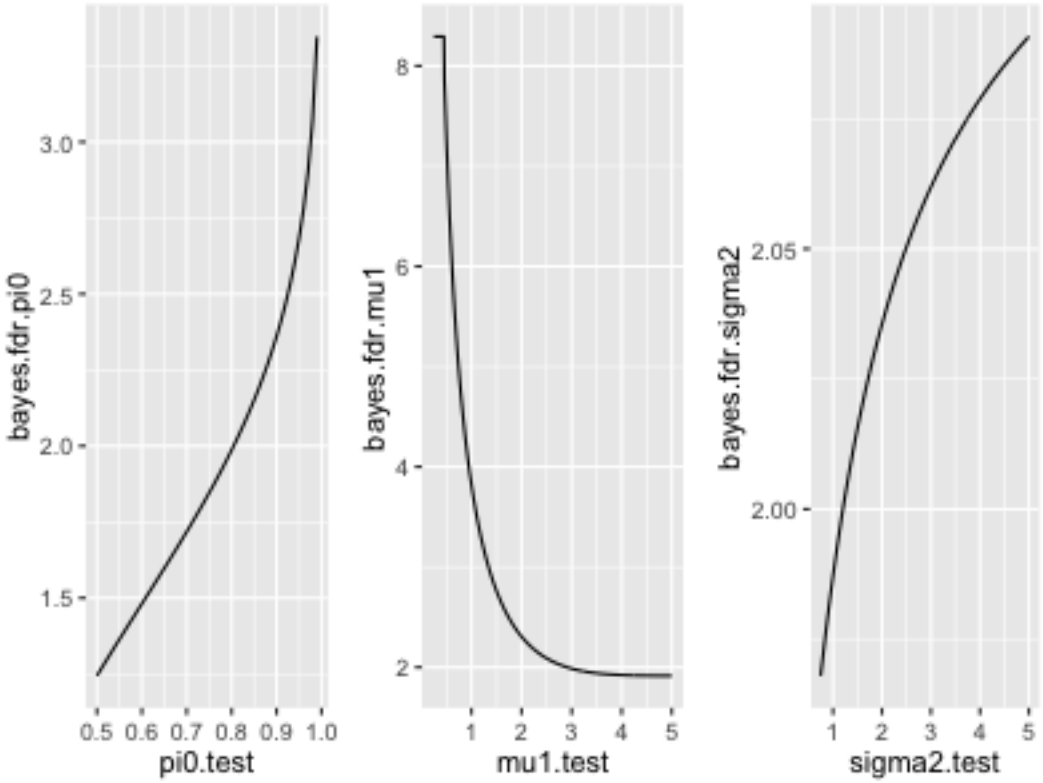

Figure 4: Plot of $\hat{x}$ while varying different simulation parameters. Simulation parameters varied are specified on the x-axis.

From the experiments above we have established that we are the most conservative when the following conditions hold:

1. π_0 is high (closer to 1)
2. μ_1 is low (closer to the null value)
3. σ_1^2 is high (the signal distribution is made more variable)

From these conclusions on the behavior of Bayesian FDR, we are able to get conservative bounds on the parameters that we estimate. We take the lower bound of the 95% credible interval for μ_1, and the upper bound of the 95% credible interval for π_0 and σ_1^2. The resulting values of $\hat{x}$ using these bounds are shown in *Figure 5*. It is promising to see that these estimates are higher than the value of $\hat{x}$ than if we knew the parameters of the mixture model beforehand.

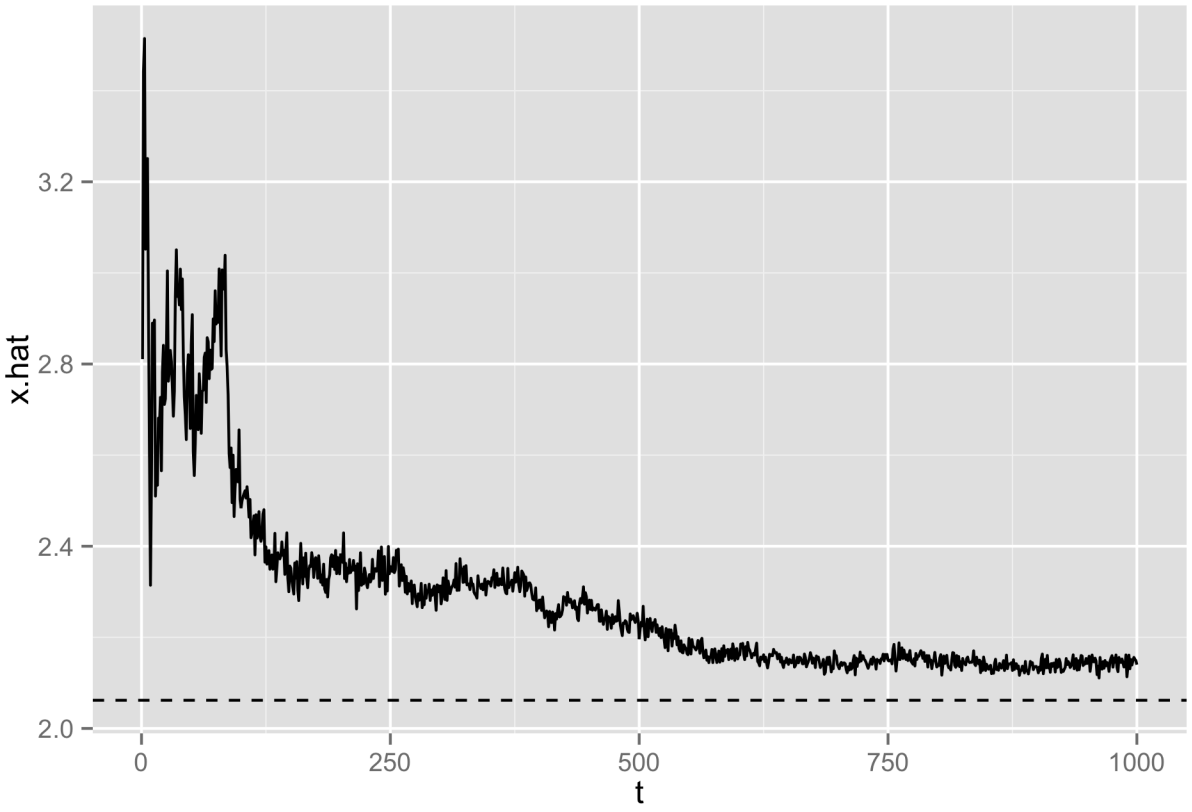

Figure 5: Plot of $\hat{x}$ with respect to time. We take the lower bound of the 95% credible interval for μ_1 and the upper bound for both π_0 and σ_1^2. The dashed line represents the $\hat{x}$ value using the true parameters for the simulation.

Empircal Estimates of FDP and Power

In order to compare our method against α-investing and LBOND/LBORD, we ran all of the methods on our simulation data and plotted the false discovery proportion (FDP) and the power. The FDR is the expected value of the FDP, and the power is the proportion of the true signals that we are able to discover.

As we see in *Figure 6* we are able to see that the most conservative methods (α-investing, and LBOND) do not have any false discoveries. LBORD has a few false discoveries, which are likely a result of resetting the thresholds after making a discovery. Our sequential Gibbs

sampling method has considerably more false discoveries than LBORD but still lies well below the global $\alpha = 0.10$ threshold.

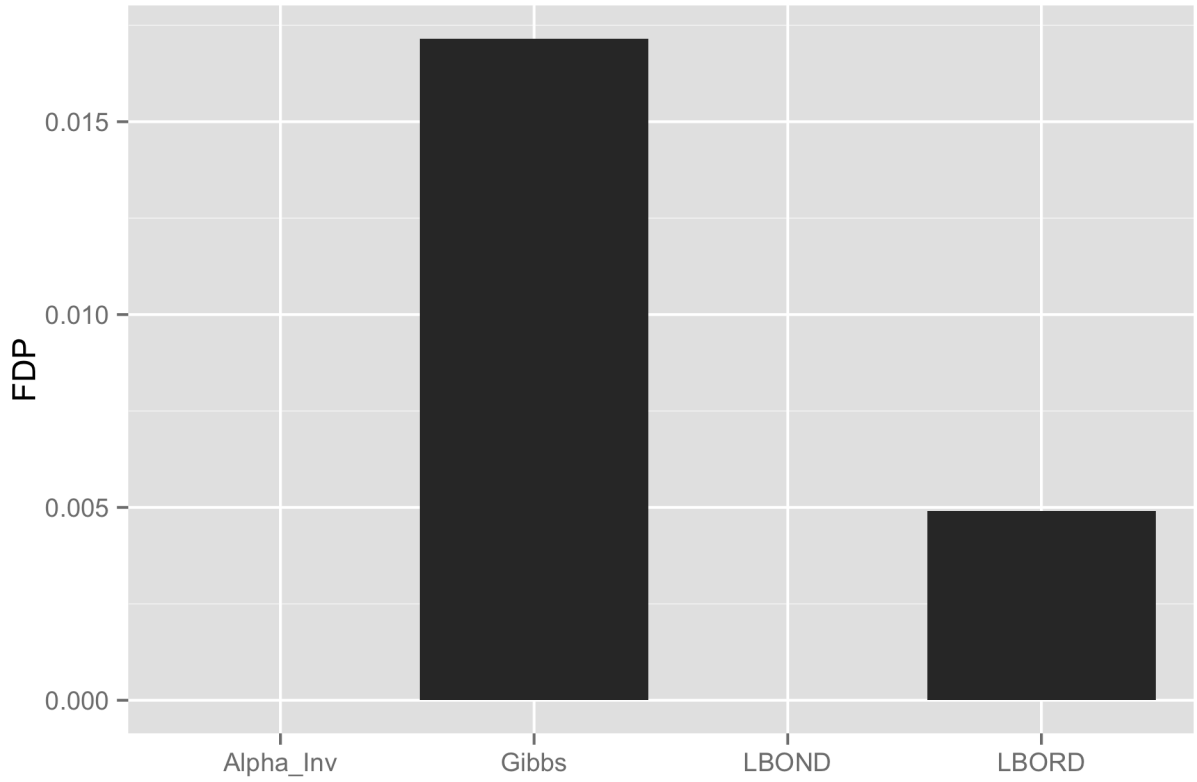

Figure 6: Empirical estimates of FDP in simulation data

Where our method is an improvement over existing methods is when comparing the power to the other online hypothesis testing methods. We see that the most conservative methods in terms of FDP are also the least powerful methods. LBORD has a considerable amount of power (~ 0.42) while retaining a lower FDP than our method, but our sequential Gibbs method has considerably higher power than LBORD (~ 0.78). The relative power between the different online testing methods can be seen in *Figure 7*.

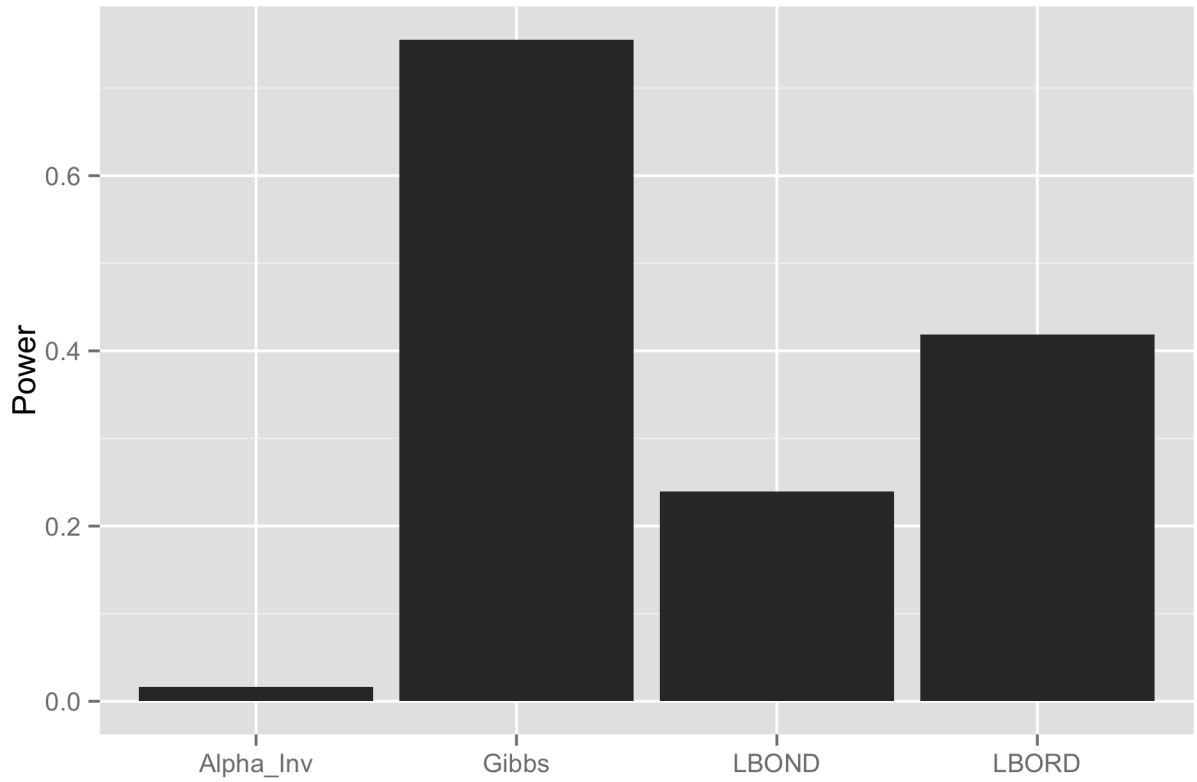

Figure 7: Empirical estimates of Power in simulation data

Conclusions

From our initial simulation study, we have shown that our method for sequential Gibbs sampling performs quite well under our simulations. However we would have to perform additional simulations in order to evaluate the performance of our new method under a variety of conditions.

It should be noted that there is no absolute guarantee that the bounds of the credible interval will always lie above and below the true parameter in question. In fact this issue may present itself more readily under different simulation parameters (i.e. under a larger proportion of nulls or a weaker signal mean)

With these caveats we propose a number of extensions and future directions for the development of our method:

1. Explore the impact of correlation within the signals. It is known that these correlations adversely affect the performance α-investing and LBOND/LBORD in terms of their

FDR control. These correlations could be set in such a way that the statistics are drawn from a Markov chain such that the stationary distribution is identical to the mixture distribution, or that there is simply a block of time during which many of the signals occur.

2. Running the Gibbs sampler "sparsely". We currently estimate the model at every time step, which means that we are performing a greater number of operations as the amount of data is increasing. We could potentially relax the invariant that we must re-estimate the model at each time-point, and instead estimate the model every k timepoints. We could potentially fit some decreasing function like LBOND/LBORD in between the parameter estimation in order to maintain proper FDR control.

3. Explore different model assumptions for the distribution of test statistics. While a mixture of normal distributions is mathematically convenient, it may not accurately represent the underlying distribution of the test statistics. One advantage of heuristic-based methods such as α-investing and LBOND/LBORD is that there are essentially no distributional requirements on the test statistics themselves, but rather that the p-values from the null distribution follow the Uniform distribution. Our method may be adaptable to different assumptions for the distribution of test statistics as well.

Code for Simulation and Replicability

We have placed our simulation code and reports generated in the following GitHub repository: https://github.com/abiddanda/online-bayesian-fdr

References

Aharoni, Ehud, and Saharon Rosset. 2014. "Generalized α-Investing: Definitions, Optimality Results and Application to Public Databases." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76 (4). Wiley Online Library: 771–94.

Efron, Bradley, and Robert Tibshirani. 2002. "Empirical Bayes Methods and False Discovery Rates for Microarrays." *Genetic Epidemiology* 23 (1). Wiley Online Library: 70–86.

Foster, Dean, and Robert Stine. 2007. "Alpha-Investing: A Procedure for Sequential Control of Expected False Discoveries." *Preprint*.

Javanmard, Adel, and Andrea Montanari. 2015. "On Online Control of False Discovery Rate." *ArXiv Preprint ArXiv:1502.06197*.

Kohavi, Ron, Roger Longbotham, Dan Sommerfield, and Randal M Henne. 2009. "Controlled Experiments on the Web: Survey and Practical Guide." *Data Mining and Knowledge Discovery* 18 (1). Springer: 140–81.