STAT 30850 Project Report

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

March 4, 2016

Contents

1. We have explicitly written out a Gibbs Sampler to estimate the mixture model (see description below). Specifically we estimate the mean and variance of the signals and the proportion of nulls as we are assuming we know the parameters of the null component.
2. We have implemented simulations as well as the Gibbs sampler and ran preliminary tests to explore how well the model is estimated, how frequently to estimate the model, and strategies for increasing performance where little data has been seen. We also compute BayesFDR each iteration using the estimated mixture model from our sampler.
3. We outline further areas we are currently planning to explore.

1. Gibbs Sampling Scheme

Let us define the following variables :

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

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 of the model conditioned on the latent indicators as:

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

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:

$$\pi_0 \mid X, Z = 0 \sim Beta(\alpha + n_0, \beta + n_1)$$
$$\phi_1 \mid X, Z \sim Gamma(\frac{a+n_1}{2}, b + \sum_{t:z_t=1}(x_t-\mu_1)^2)$$
$$\mu_1 \mid X, Z, \phi_1 \sim Normal(\frac{\alpha^*\mu^* + n_1 + \bar{x_1}}{\alpha^* + n_1}, \frac{1}{(\alpha^* + n_1)\phi_1})$$

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

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

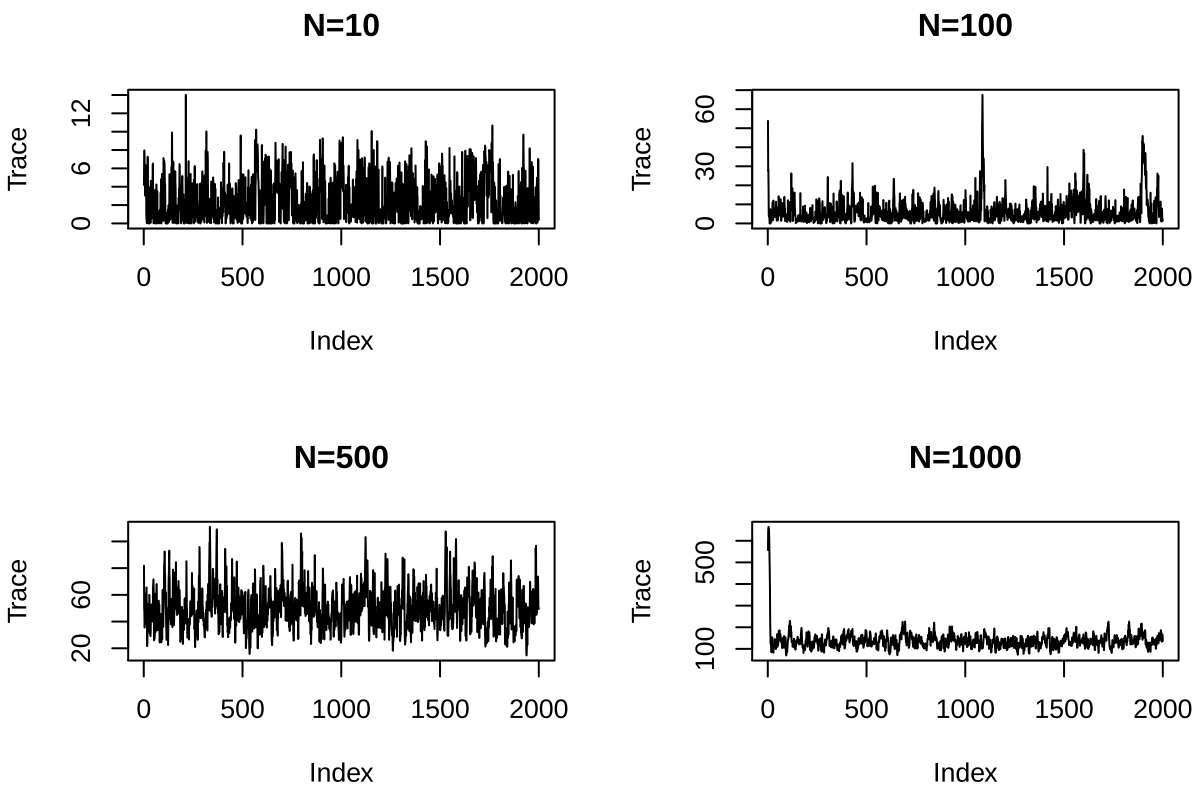
We have implemented a Gibbs sampler to sample from the posterior distributions of the parameters:

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

2. Simulations

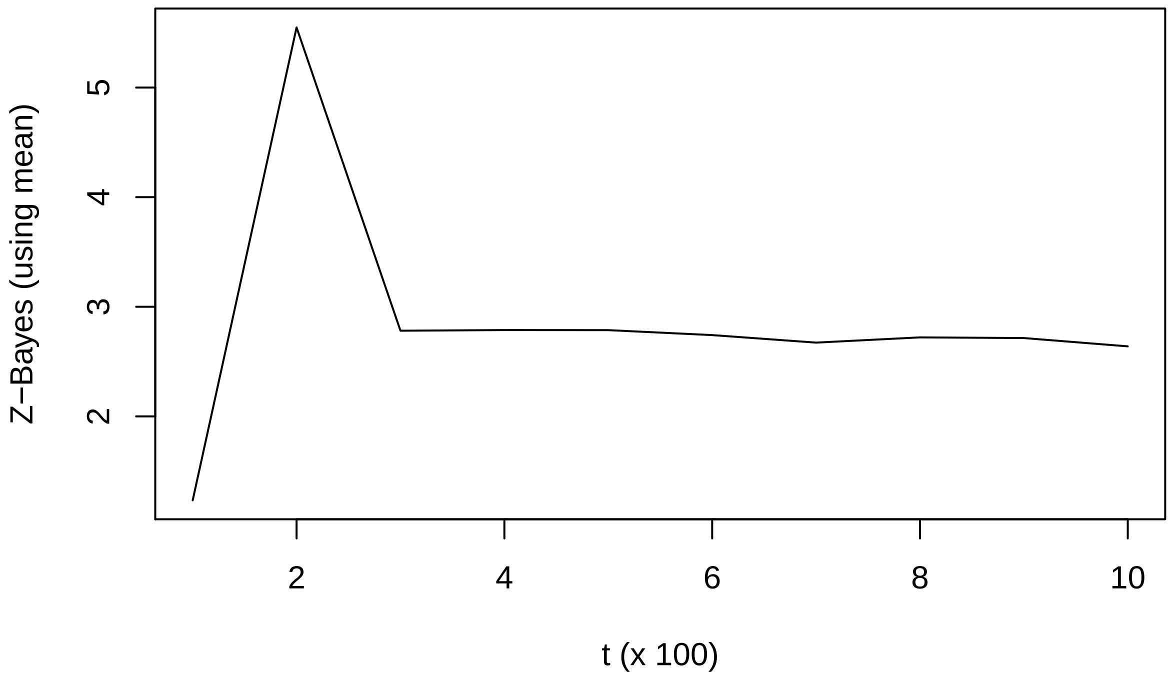
We have implemented a framework such that we stream data from a known simulated mixture model and estimate this model using Gibbs sampling. As seen below we can successfully estimate this model after we've observed a reasonable amount of data.

Figure 1



Trace plot of Gibbs sampler. Trace on the y-axis is the negative log-likelihood of the the latent Z which depends on all the parameters of the mixture model. We see that the sampler does not mix well early in the data-stream i.e. with little data.

Figure 2



Plot of BayesianFDR Z-score critical value over time. We see strange results in the beginning of the data-stream due to poor estimation of the the mixture model. As time proceeds reasonable critical values are computed.

3. Future Plans / Exploration

- Explore what summary of the posterior distribution is most appropriate and conservative to use for the time-wise BayesFDR i.e. posterior mean, upper credible interval.
- Explore how well this approach controls FDR and retains power.
- Come up with strategies in the beginning of the data stream to be more conservative and have less confidence in our estimated model.
- Explore how often we should estimate the model (interval of sampling)
- Explore the effect of temporal correlation in the signals and its effects on parameter estimation and FDR.
- Implement the Gibbs sampler in C++ for efficiency