

# Bayesian Nonparametric Multiple Testing

William Cipolli III  
University of South Carolina  
Department of Statistics

Timothy Hanson  
University of South Carolina  
Department of Statistics

Alexander C. McLain  
University of South Carolina  
Department of Epidemiology and Biostatistics

## Abstract

Multiple testing, or multiplicity problems often require testing several means with the assumption that we will reject infrequently, as motivated by the need to analyze DNA microarray data. The goal is to keep the combined rate of false discoveries and non-discoveries as small as possible. We propose a discrete approximation to a Polya tree prior that enjoys fast, conjugate updating, centered at the usual normal distribution, thus generalizing Scott and Berger (2006) to a nonparametric setting. This new technique and the advantages of our approach are demonstrated using extensive simulation and data analysis accompanied by a Java web application. The numerical studies demonstrate that our new procedure shows promising FDR and estimation of key values in the mixture model with very reasonable computational speed.

Keywords: Multiplicity, Polya tree mixture, False Discovery Rate, MCMC, Deconvolution, Mixtures

## 1 Introduction

In the DNA microarray setting, multiple testing problems often require testing multiple means  $\theta_i = 0$ . The  $\theta_i$  measurements can contain a range of values, both positive and negative, and we want to detect the  $|\theta_i|$  that are large enough to reject  $\theta_i = 0$ .

We initially consider  $y_i$ ,  $i = 1, \dots, n$  such that each  $y_i$  is independently normally distributed with mean  $\theta_i$  and fixed, unknown variance  $\sigma^2$ , as considered by Scott and Berger (2006); we follow with a model that accounts for known variances  $\sigma_i^2$ , as considered by Sun and McLain (2012).

We consider two tasks while testing  $\theta_i = 0$ : the task of finding  $\theta_i$  different from, and “interestingly different” from zero. In the DNA microarray setting, often  $y_i$  are difference measurements of gene expression in two states and low differences are of little interest; scientists consider genes that show at least two-fold change in expression levels to be differentially expressed, meaning observations with large deviations from the null are sought. It is important to distinguish that this model, as stated, executes the hypothesis tests that  $\theta_i$  are different from zero but we provide adapted methodology to answer the question of “interestingly different” observations.

### 1.1 Biological Motivation

An organism’s basic make up is due to its genome, among plants, animals and humans alike. Genomes are comprised of one or many strands of DNA. Each living species, and the individual

variations within those species, are defined through the details of DNA. A single cell of a living organism contains at least one copy DNA which is organized into chromosomes; humans have twenty three pairs of chromosomes, most famously the paired X and Y sex chromosomes. The chromosomes, which make up DNA, contain regions called genes that are involved in the production of proteins; each chromosome can contain a different collection of genes.

Genes are important to medical research because they are involved in the production of proteins which make up the entirety of the organisms by controlling replication, form and mutation. When you look at a human, for instance, everything you see consists of protein: hair, skin, eyes, etc. The rate of cell production is also controlled by genes so the regeneration of skin cells and hair growth etc. are also attributed to genes. Sometimes, though rarely, a genetic mutation can cause a disease like Progeria which causes accelerated aging as illustrated in the 1996 film *Jack* starring the late Robin Williams.

The influence that genes have on protein production, and consequently an organism's properties, make their influence important. Before DNA microarray data was available, traditional methodologies only provided ways to study one or a few genes at a time; now widespread availability makes it possible to explore all the genes in a single experiment.

DNA microarray data displays sets of microscopic spots of DNA laid out on a piece of glass; together this is referred to as a DNA chip. DNA chips come in two flavors, cDNA or oligonucleotides. cDNA requires that the DNA chip have full-length transcripts printed onto the glass and oligonucleotides are chemically synthesized on the glass and then exposed to probes which extract information from a particular cell during different stages. Scientists usually prefer oligonucleotides because these nucleotides can identify which genes are active at certain stages of development, or under certain environmental conditions. The goal is often to define which genes, and subsequently which proteins, are active or inactive during these stages or environments.

Efron et al. (2001) and Do et al. (2005) considered a prevalent example of multiple testing when analyzing DNA microarray data. The problem explored was how certain treatments, environment changes or disease affects gene expression. The DNA microarray data of Efron et al. (2001) consists of 6,810 genes exposed to eight conditions. Even if all the genes were not activated, meaning we should fail to reject each  $\theta_i = 0$ , setting up the hypothesis tests with a significance level .05 leads to making a Type I error roughly five percent of the time; if this were the case when testing 6,810 genes we would expect to make Type I error about 340 times.

## 1.2 Motivations Beyond Biology

Research in multiple testing has been heavily motivated by DNA microarray data. The need for developing methods that accurately control for multiple testing should increase as the world more heavily depends on big data solutions in both science and business, which are becoming more prevalent. Multiple testing can be useful in multiple comparison problems including brain imaging (Chumbley and Friston, 2009), financial analysis (Baj, 2008), market research (Blomquist, 2014) and many other data heavy sectors that exist today as well as future data-based technologies.

Sun and McLain (2012) provide another prominent example affected by heteroscedasticity; educational survey data from the Adequate Yearly Progress (AYP) study on the academic performances of students across different social and fiscal demographics. These data are reanalyzed in Section 7.

## 2 Models

We have data  $\mathbf{y} = (y_1, \dots, y_n)'$  with mean vector  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)'$ . Consider the following Gaussian mixture model on the responses

$$y_i | \boldsymbol{\theta}, \sigma^2 \stackrel{ind.}{\sim} N(\theta_i, \sigma^2), \quad \theta_i | w, G \stackrel{ind.}{\sim} wg(\theta) + (1 - w)\delta_0, \quad (1)$$

where  $G(\cdot)$  is a cumulative distribution function with corresponding density  $g(\cdot)$ , and  $w$  is the mass parameter in the mixture model denoting the proportion of non-null observations, and  $\delta_x$  is the Dirac measure at  $x$ . Scott and Berger (2006) consider a parametric case where  $G$  is Gaussian with mean zero and standard deviation  $v$ , i.e.  $G = N(0, v^2)$ . We generalize this model by assigning  $G$  a nonparametric finite Polya tree prior with  $J$  levels that is centered at a  $N(0, v^2)$  distribution. That is,

$$G | c, v^2 \sim PT_J(c, w, N(0, v^2)). \quad (2)$$

where  $PT_J$  denotes a Polya tree of  $J$  levels, and  $c$  is a parameter that controls how “close”  $G$  is to the centering distribution  $N(0, v^2)$ . The prior on  $w$ , the amount of mass assigned to  $g(\theta)$ , is beta

$$w \sim \text{beta}(a_w, b_w). \quad (3)$$

For this particular prior Scott and Berger (2006) set  $a_w = \alpha + 1$  and  $b_w = 1$  with  $\alpha = 1$ . In many experiments, however, experts will provide information on  $w$ , the proportion of non-null observations. In this case, consider taking  $a_w = \hat{p}m$  and  $b_w = (1 - \hat{p})m$ , where  $\hat{p} \in (0, 1)$  and  $m > 0$  so that  $E(w) = \hat{p}$ , matching the information provided by the expert, and choosing  $m$  so that  $\text{var}(w) = \frac{\hat{p}(1-\hat{p})}{m+1}$  matches prior belief.

Deciphering which observations are “interestingly different” from zero has much to do with the prior on  $w$ , the amount of mass assigned to  $g(\theta)$  or the proportion of non-null observations. By altering the prior on  $w$  we allow the user to indicate how many of the observations should be rejected, allowing the tweaking of  $w$  to be interpreted as the proportion of “interestingly different” observations instead of just observations with non-zero means.

Consider the case where we find  $w = w_0$ . When observations that are “interestingly different” than zero are desired, we can then try using prior information  $\hat{p} < w_0$  and  $m$  so that  $E(w) = \hat{p}$  and  $\text{var}(w)$  matches belief. As  $\hat{p}$  decreases we will see less, but “more interesting” rejections. We demonstrate this in the simulations and data analyses in Section 6 and Section 7.

Finally, the variance components  $\sigma^2$  and  $v^2$  have independent inverse gamma priors

$$\sigma^{-2} \sim \Gamma(a_\sigma, b_\sigma) \text{ independent of } v^{-2} \sim \Gamma(a_v, b_v). \quad (4)$$

Scott and Berger (2006) consider the improper prior  $p(\sigma^2, v^2) \propto (v^2 + \sigma^2)^{-2}$ .

The Polya tree prior was initially summarized by Ferguson (1974), and further developed by Lavine (1992, 1994), and Mauldin et al. (1992). Hanson (2006) discusses inference for mixtures of finite Polya trees, which smooth out the effect of the partition on posterior inference. Briefly, the prior (2) on  $G$  adds to  $N(0, v^2)$ . The  $2^J - 1$  conditional probabilities that refine  $G$ ’s shape are

$$Y_{j,k} | c \stackrel{ind.}{\sim} \text{beta}(cj^2, cj^2),$$

where  $j = 1, \dots, J$ , and  $k$  are the odd numbers from 1 to  $2^j - 1$  at any level  $j$ . For any  $Y_{j,k}$  where  $k$  is odd, let  $Y_{j,k+1} = 1 - Y_{j,k}$ . Define  $\mathcal{Y} = \{Y_{jk} : j = 1, \dots, J; k = 1, \dots, 2^j\}$ . A Polya tree parameter is the conditional probability  $Y_{j,k} = G\{B_v(j, k) | B_v(j - 1, \lceil k/2 \rceil)\}$  where  $B_v(j, k) = (v\Phi^{-1}((k - 1)/2^j), v\Phi^{-1}(k/2^j))$ , is the interval for the partition  $k$  on level  $j$ . Note that  $B_v(j, 1), \dots, B_v(j, 2^j)$

partitions  $\mathbb{R}$  up to a set of measure zero and for any measurable  $A \subset \mathbb{R}$ ,  $E\{G(A)\} = \int_A \phi(t|0, v^2)dt$  where  $\phi(\cdot|\mu, \sigma^2)$  is the density of a Gaussian random variable with mean  $\mu$  and standard deviation  $\sigma$ . The probability of being in set  $k$  at level  $J$  is:

$$p_{\mathcal{Y}}(k) = \prod_{j=1}^J Y_{j, \lceil k2^{j-J} \rceil}, \quad k = 1, \dots, 2^J.$$

with  $\lceil \cdot \rceil$  the usual ceiling function. For example, for  $J = 5$  we can find  $p_{\mathcal{Y}}(11)$  as highlighted in Figure 1:

$$\begin{aligned} p_{\mathcal{Y}}(k = 11) &= \prod_{j=1}^{J=5} Y_{j, \lceil (11)2^{j-5} \rceil} \\ &= Y_{1, \lceil (11)2^{1-5} \rceil} Y_{2, \lceil (11)2^{2-5} \rceil} Y_{3, \lceil (11)2^{3-5} \rceil} Y_{4, \lceil (11)2^{4-5} \rceil} Y_{5, \lceil 112^{5-5} \rceil} \\ &= Y_{1,1} Y_{2,2} Y_{3,3} Y_{4,6} Y_{5,11}. \end{aligned} \tag{5}$$

Although the measure  $G$  is discrete, the usual density estimate from a mixture of Polya trees can be used to smooth  $G$  and give an idea of how mass is spread out. Another option is to simply plot the location and height of the point masses. We opt for the former using formula (6) in Hanson (2006). For  $G \sim PT_J(c, \rho, N(0, v^2))$ , the density  $g(\theta) = g(\theta|\mathcal{Y}, v^2)$  of  $G(\theta)$  given  $\mathcal{Y}$  and  $v$  has the following form:

$$g(\theta|\mathcal{Y}, v^2) = 2^J p\{k_v(\theta)\} \phi(\theta|0, v^2) = 2^J \phi(\theta|0, v^2) \sum_{k=1}^{2^J} I\{k_v(\theta) = k\} p_{\mathcal{Y}}(k),$$

where  $2^J$  gives the number of partitions in the last level of the Polya tree;  $k_v(\theta) = \lceil 2^J \Phi(\theta/v) \rceil$ , gives which set  $k$ , at level  $J$ ,  $\theta$  is in.

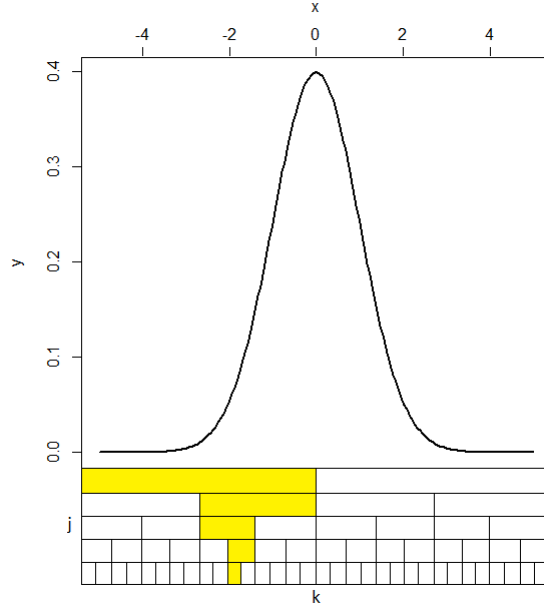


Figure 1: Graph of normal pdf with Polya tree partitions(left) with example highlights(right)

## 2.1 Direct inference via adaptive MCMC

Following Scott and Berger (2006) we first consider marginalized inference. Marginalizing over  $\theta$ , each datum arises independently from the density

$$f(y_i|\sigma^2, w, G) = w \int_{\mathbb{R}} g(\theta) \phi(y_i|\theta, \sigma^2) d\theta + (1-w) \phi(y_i|0, \sigma^2).$$

For  $a < b$  note the identity

$$\int_a^b \phi(y|\theta, \sigma^2) \phi(\theta|0, v^2) d\theta = \phi(y|0, \sigma^2 + v^2) \left[ \Phi \left\{ \frac{b(\sigma^2 + v^2) - v^2 y}{\frac{\sigma v}{\sqrt{\sigma^2 + v^2}}} \right\} - \Phi \left\{ \frac{a(\sigma^2 + v^2) - v^2 y}{\frac{\sigma v}{\sqrt{\sigma^2 + v^2}}} \right\} \right].$$

This leads to the density of  $y$  given  $v$ ,  $\sigma$ , and  $\mathcal{Y}$  as

$$\begin{aligned} m(y|v, \sigma, \mathcal{Y}, w) &= w \int_{\mathbb{R}} g(\theta|\mathcal{Y}, v) \phi(y|\theta, \sigma^2) d\theta + (1-w) \phi(y_i|0, \sigma^2) \\ &= w \int_{\mathbb{R}} 2^J \phi(\theta|0, v^2) \left[ \sum_{k=1}^{2^J} I\{k_v(\theta) = k\} p_{\mathcal{Y}}(k) \right] \phi(y|\theta, \sigma^2) d\theta + (1-w) \phi(y_i|0, \sigma^2) \\ &= w \int_{\mathbb{R}} 2^J \left[ \sum_{k=1}^{2^J} I\{k_v(\theta) = k\} p_{\mathcal{Y}}(k) \right] [\phi(y|\theta, \sigma^2) \phi(\theta|0, v^2)] d\theta + (1-w) \phi(y_i|0, \sigma^2) \\ &= w 2^J \phi(y|0, \sigma^2 + v^2) \sum_{k=1}^{2^J} p_{\mathcal{Y}}(k) \Delta(y, k|\sigma, v) + (1-w) \phi(y_i|0, \sigma^2). \end{aligned}$$

where

$$\Delta(y, k|v, \sigma) = \Phi \left\{ \frac{\Phi^{-1} \left( \frac{k}{2^J} \right) (\sigma^2 + v^2) - y}{\frac{\sigma}{\sqrt{\sigma^2 + v^2}}} \right\} - \Phi \left\{ \frac{\Phi^{-1} \left( \frac{k-1}{2^J} \right) (\sigma^2 + v^2) - y}{\frac{\sigma}{\sqrt{\sigma^2 + v^2}}} \right\}.$$

The unnormalized posterior density is then

$$\begin{aligned} p(v, \sigma, c, \mathcal{Y}, w) &\propto \left[ \prod_{i=1}^n m(y_i|v, \sigma, \mathcal{Y}, w) \right] p(v) p(\sigma) p(c) \\ &\quad \times \left[ \prod_{j=1}^J \prod_{k=1}^{2^{j-1}} \text{beta}(Y_{j,2k-1} | c j^2, c j^2) \right] \text{beta}(w | a_w, b_w). \end{aligned}$$

The dimension of the posterior parameter vector is  $2^J + 3$ . Adaptive Metropolis-Hastings (Haario et al., 2001) can be used here to obtain posterior inference.

## 3 Discrete approximation to simplify posterior updating

### 3.1 Discrete approximation

To simplify the computational complexity, we consider a discrete approximation to the Polya tree. Define  $G$  to be the finite discrete measure

$$G(\cdot) = \sum_{k=1}^{2^J} p_{\mathcal{Y}}(k) \delta_{\theta_k}(\cdot), \quad \theta_k = v \Phi^{-1} \left( \frac{k - 0.5}{2^J} \right) \stackrel{\text{def}}{=} v t_k. \quad (6)$$

Note that as  $J$  gets large, the intervals  $B_v(J, k)$  get smaller, except in the tails, and  $g(\cdot)$  varies less over the intervals. Here  $g(\cdot)$  follows  $N(0, v^2)$  over each interval  $k$ , and can be approximated with just one “representative” point, the mid-quantile,  $\theta_k$ , plus the associated probability  $p_{\mathcal{Y}}(k)$  of the interval under  $G$ . This leads to the density of  $y_i$  given  $\sigma^2, w, v, \mathcal{Y}$  as

$$\begin{aligned} f(y_i|\sigma^2, w, v, \mathcal{Y}) &= w \int_{\mathbb{R}} \phi(y_i|\theta, \sigma^2) G(d\theta) + (1-w)\phi(y_i|0, \sigma^2) \\ &= w \sum_{k=1}^{2^J} p_{\mathcal{Y}}(k) \phi(y_i|\theta_k, \sigma^2) + (1-w)\phi(y_i|0, \sigma^2). \end{aligned}$$

Note that we can marginalize over  $\sigma^2$  with

$$\int_0^\infty \phi(y_i|\theta, \sigma^2) \Gamma(\sigma^{-2}|a, b) d\sigma^{-2} = \frac{b^a}{\sqrt{2}\beta(a, 1/2)} \left[ b + \left( \frac{y_i - \theta}{\sqrt{2}} \right)^2 \right]^{-1/2-a}.$$

This can form the basis of inference using a “black box” sampler, for example an adaptive Metropolis-Hastings proposal (Haario et al., 2001).

### 3.2 Gibbs sampling through data augmentation

To simplify computation, component membership indicators are introduced. Let  $z_i = j$  iff  $y_i \sim N(\theta_j, \sigma^2)$ , where  $\theta_0 = 0$  and  $j = 0, 1, \dots, 2^J$ . Then

$$P(z_i = k|v, \sigma^2, w, \mathcal{Y}) \propto \begin{cases} \phi(y_i|0, \sigma^2)(1-w) & k = 0 \\ \phi(y_i|t_k v, \sigma^2) w p_{\mathcal{Y}}(k) & k > 0 \end{cases}.$$

$$f(v|\sigma^2, \mathbf{z}) \propto f(v) \prod_{i: z_i \neq 0} \exp\{-0.5\sigma^{-2}(y_i - t_{z_i} v)^2\},$$

so then

$$f(v|\sigma^2, \mathbf{z}) \propto N\left(v \mid \frac{\sum_{i=1}^n t_{z_i} y_i}{\sum_{i=1}^n t_{z_i}^2}, \frac{\sigma^2}{\sum_{i=1}^n t_{z_i}^2}\right) f(v) = N\left(v \mid \frac{\sum_{i: z_i > 0} t_{z_i} y_i}{\sum_{i: z_i > 0} t_{z_i}^2}, \frac{\sigma^2}{\sum_{i: z_i > 0} t_{z_i}^2}\right) f(v).$$

Here,  $v^*$  is sampled from the Gaussian full conditional (under a flat prior) above and accepted with M-H probability  $1 \wedge f(v^*, \sigma^2)/f(v^2, \sigma^2)$ , e.g.  $f(v^2, \sigma^2) = \frac{1}{v^2 + \sigma^2}$  is the improper prior suggested by Scott and Berger (2006).

This is explicitly given to emphasize that the conditional distribution depends only on those observations for which  $z_i > 0$ , i.e. those observations deemed to be from the non-null distribution. During the Gibbs sampler one can sample and keep track of the  $z_i$  while simultaneously updating  $v$  and the means  $t_k v$ .

Note the non-identifiability of the model: different values of  $\mathcal{Y}$  and  $v$  can give the same likelihood:  $v \rightarrow -v$  and  $p_{\mathcal{Y}}(k) \rightarrow p_{\mathcal{Y}}(2^J - j + 1)$ . Thus  $v > 0$  is needed and maintains interpretation of  $v$  as a scale parameter.

Consider the full conditional distributions are

$$w|\mathbf{z} \sim \text{beta}\left(a_w + \sum_{i=1}^n I\{z_i > 0\}, b_w + \sum_{i=1}^n I\{z_i = 0\}\right),$$

with  $a_w = \alpha + 1$  and  $b_w = 1$ , and

$$\sigma^{-2}|\mathbf{z}, v \sim \Gamma \left( a_\sigma + 0.5n, b_\sigma + 0.5 \sum_{i=1}^n (y_i - vt_{z_i})^2 \right),$$

where  $t_0 = 0$ . Let

$$n(J, k) = \sum_{i=1}^n I\{z_i = k\}, \quad n(j-1, k) = n(j, 2k-1) + n(j, 2k), \quad j = 2, \dots, J,$$

then

$$Y_{j, 2k-1} \sim \text{beta}(c\rho(j) + n(j, 2k-1), c\rho(j) + n(j, 2k)), \quad k = 1, \dots, 2^{j-1}, \quad j = 1, \dots, J.$$

Note that  $Y_{1,1} = 0.5$  and is not sampled.

Finally,

$$p_{\mathcal{Y}}(k) = \prod_{j=1}^J Y_{j, \lceil k/2^{j-1} \rceil}, \quad k = 1, \dots, 2^J.$$

The probabilities  $\mathcal{Y}$  are “unhinged” from the location of the sets, or points  $\theta_k$ . This makes MCMC sampling easy.

## 4 Deconvolution with known variances

We now consider the situation where the  $y_i$  are observed with known heteroscedastic error  $\sigma_i^2$ :

$$y_i | \boldsymbol{\theta}, \sigma^2 \stackrel{\text{ind.}}{\sim} N(\theta_i, \sigma_i^2),$$

as considered by Sun and McLain (2012). For this model,  $\sigma^2$  no longer needs to be updated. Component membership is updated via

$$P(z_i = k | v, \sigma_i^2, w, \mathcal{Y}) \propto \begin{cases} \phi(y_i | 0, \sigma_i^2)(1-w) & k = 0 \\ \phi(y_i | t_k v, \sigma_i^2) w p_{\mathcal{Y}}(k) & k > 0 \end{cases}.$$

The full conditional for  $v$  is now updated to

$$f(v | \boldsymbol{\sigma}, \mathbf{z}) \propto f(v) \prod_{i: z_i \neq 0} \exp\{-0.5\sigma_i^{-2}(y_i - t_{z_i}v)^2\},$$

where

$$\begin{aligned} f(v | \boldsymbol{\sigma}, \mathbf{z}) &\propto N \left( v \left| \frac{\sum_{i=1}^n t_{z_i} y_i / \sigma_i^2}{\sum_{i=1}^n t_{z_i}^2 / \sigma_i^2}, \frac{1}{\sum_{i=1}^n t_{z_i}^2 / \sigma_i^2} \right| \right) f(v) \\ &= N \left( v \left| \frac{\sum_{i: z_i > 0} t_{z_i} y_i / \sigma_i^2}{\sum_{i: z_i > 0} t_{z_i}^2 / \sigma_i^2}, \frac{1}{\sum_{i: z_i > 0} t_{z_i}^2 / \sigma_i^2} \right| \right) f(v). \end{aligned}$$

In what follows we simply take  $f(v) \propto I_{(0,u)}(v)$  for some large  $u > 0$ .

It should be noted that setting  $c$  and  $J$  large provides a close approximation to Scott and Berger (2006) but also gives a close approximation to that model extended to this known variance case.

## 5 Error rates

With single hypothesis tests the rejection threshold is moved to control Type I error. In the case of multiplicity, there may be hundreds, thousands or possibly millions of hypothesis tests each having their own Type I and Type II errors. These errors need to be combined to talk about the overall error, i.e. specificity or sensitivity of multiple hypothesis tests.

|                      | $H_0$ True             | $H_0$ False        | Total                |
|----------------------|------------------------|--------------------|----------------------|
| Reject $H_0$         | Type I errors          | Correct Rejections | Total Rejections     |
| Fail to Reject $H_0$ | Correct Non-Rejections | Type II errors     | Total Non-Rejections |
| Total                | Total True             | Total False        | Total Tests          |

Table 1: Errors for many hypothesis tests

Family wise error rate (FWER) methods; i.e. Sidak (1967), Tukey (1949), and Hochberg (1988), look at the probability of making one or more Type I errors. The goal is to minimize this probability, i.e. reduce the probability of seeing any Type I error.

$$\begin{aligned}
 FWER &= P(\text{False Discoveries} \geq 1) \\
 &= 1 - P(\text{False Discoveries} = 0) \leq \alpha
 \end{aligned}$$

The problem here, in line with one single hypothesis test, is that minimizing this probability leads to an increase in Type II error, false non-discovery.

The false discovery rate, FDR, introduced by Benjamini and Hochberg (1995) has become the standard for error measurement in multiple testing problems; it provides a way to control Type I error like FWER but does a better job of balancing the Type II error by allowing more Type I error. It should also seem more logical to consider the rate or proportion of errors instead of whether or not any errors were made; five false discoveries in ten hypothesis tests is much worse than the same number out of one hundred tests. Müller et al. (2007) suggested a variation of FDR from a conditional Bayesian perspective, but we continue with the traditional measurement.

FDR starts with the calculation of the false discovery proportion, FDP, as the proportion of false discoveries among all discoveries, closely aligned with Type I error. FDR is then the expected value of FDP. Similarly, the false non-discovery rate, FNR, starts with the calculation of the false non-discovery proportion, FNP, as the proportion of false non-discoveries among all non-discoveries, aligned with Type II error. FNR is the expected value of FNP.

The tests that we are interested in are  $H_i : \theta_i = 0$ ;  $i = 1, \dots, n$ . Let  $\gamma_i = (\gamma_1, \dots, \gamma_n)$  with  $\gamma_i = I(\theta_i = 0)$ . Let  $\tau = (\tau_1, \dots, \tau_n)$  be the collection of test statistics  $\tau_i = P(z_i = 0 | \mathcal{Y})$ , the posterior probability that we fail to reject  $H_i$ , with test threshold values of  $T = (T_1, \dots, T_n)$ , with default

$$\begin{aligned}
 T_{i1} &= \frac{k \int_0^\infty |\frac{\theta_i}{v}| \pi(\theta_i | \gamma_i = 0, \mathbf{y}) d\theta_i}{1 + k \int_0^\infty |\frac{\theta_i}{v}| \pi(\theta_i | \gamma_i = 0, \mathbf{y}) d\theta_i} \\
 &= \frac{kE(|\frac{\theta_i}{v}|)}{1 + kE(|\frac{\theta_i}{v}|)} \\
 &= \frac{kE(|t_k|)}{1 + kE(|t_k|)},
 \end{aligned}$$

which is based on the threshold proposed by Scott and Berger (2006). The relative cost of making a false non-discovery compared to a false discovery is denoted by  $k$ . All else equal,  $T_{i1}$  is increasing

in  $k$  indicating that if the cost of a false non-discovery is higher relative to a false discovery, larger  $k$  is desirable.

In many cases, large observations cause the threshold of Scott and Berger (2006) to approach one, regardless of the observations' variances. This causes observations with large  $E(|\theta_i|)$  to be rejected overall even if they were rejected just once during Gibbs sampling. This is an important consideration because real world observations can be quite large; the carcinoma data from Notterman et al. (2001), considered in the data analysis of Section 7, has difference values over 1,300 for example. To account for these cases scaling  $\theta_i$  by the variance term  $v$ , is required.

Alternatively, we can use

$$\begin{aligned} T_{i2} &= \frac{k \int_0^\infty |\theta_i| \pi(\theta_i | \gamma_i = 0, \mathbf{y}) d\theta_i}{1 + k \int_0^\infty |\theta_i| \pi(\theta_i | \gamma_i = 0, \mathbf{y}) d\theta_i} \\ &= \frac{kE(|\theta_i|)}{1 + kE(|\theta_i|)} \\ &= \frac{kE(|t_k v|)}{1 + kE(|t_k v|)}. \end{aligned}$$

The motivation for the unscaled threshold,  $T_{i2}$ , of Scott and Berger (2006) was to create a threshold such that observations with posterior means close to zero have lower thresholds than those with posterior means far from zero. The threshold  $T_{i2}$  is increasing in  $|\theta_i|$ , which causes the issues with large observations mentioned above. This threshold, however, is very comparable to  $T_{i1}$  when considering observations of magnitude  $10^1$  and performs better in the case when considering the observations of magnitude  $10^{-1}$  or smaller as the threshold  $T_{i1}$  is inflated for small  $\theta_i$ .

Finally, this gives the decision functions  $\delta_{i1} = I(\gamma_i < T_{i1})$  and  $\delta_{i2} = I(\gamma_i < T_{i1})$  which we toggle between depending on the magnitude of our observations.

|                      | $H_0$ True                                 | $H_0$ False                          | Total                       |
|----------------------|--------------------------------------------|--------------------------------------|-----------------------------|
| Reject $H_0$         | $V = \sum_i (1 - \gamma_i) \delta_i$       | $S = \sum_i \gamma_i \delta_i$       | $R = \sum_i \delta_i$       |
| Fail to Reject $H_0$ | $U = \sum_i (1 - \delta_i) (1 - \gamma_i)$ | $W = \sum_i (1 - \delta_i) \gamma_i$ | $A = \sum_i (1 - \delta_i)$ |
| Total                | $\sum_i (1 - \gamma_i)$                    | $\sum_i (\gamma_i)$                  | $n$                         |

Table 2: Errors Calculation for Multiple Hypothesis Tests

The number of discoveries can be written as  $R = \sum_i \delta_i$ , with the number of false discoveries  $V = \sum_i (1 - \gamma_i) \delta_i$  making  $FDP = \frac{V}{\max(R, 1)}$ . FDR is then  $E\left(\frac{V}{\max(R, 1)}\right)$ . The number of non-discoveries can be written as  $A = \sum_i (1 - \delta_i)$ , with the number of false non-discoveries  $W = \sum_i (1 - \delta_i) \gamma_i$  making  $FNP = \frac{W}{\max(A, 1)}$ . FNR is then  $E\left(\frac{W}{\max(A, 1)}\right)$ . In the same vein, the missed discovery rate, MDR, is  $E\left(\frac{W}{\max(m_0, 1)}\right)$  where  $m_0 = \sum_i \gamma_i$ , the number of  $\theta_i$  that should be rejected. The calculations should clarify the difference of FNR and MDR. FNR is the ratio of false non-discoveries among all non-discoveries whereas MDR is the ratio of false non-discoveries among all  $\theta_i$  that should be rejected.

FDR and FNR are the expected value of FDP and FNP respectively and a simple Law of Large Numbers argument allows us to approximate FDR and FNR with FDP and FNP respectively in the simulation studies in Section 6. The marginal versions are defined as  $MFDR = \frac{E(V)}{E(\max(R, 1))}$  and  $MMDR = \frac{E(W)}{E(\max(m_0, 1))}$ . Genovese and Wasserman (2002) show that MFDR and FDR are negligibly different in large problems and are argued better than their non-marginal counterparts by Storey (2002) and Wu and Cai (2007).

A more advanced procedure for error measurement is discussed by Peña et al. (2011) who consider taking the individual powers of each test, which are left out of other methods, into account for both FWER and FDR methods. Their research focuses on Neyman-Pearson Most Powerful tests of simple hypotheses with promising results that indicate individual powers are important to multiple tests. Our paper does not include a model that accounts for individual power, but it is an interesting avenue to consider in further research.

## 6 Simulations

Two scenarios are considered for simulation, one where the means are distributed  $G_1 = N(0, 2^2)$ , and another where the means follow a skewed, bimodal, median-zero mixture of two normals  $G_2 = 0.4N(-2.62, 1^2) + 0.6N(0.48, 0.5^2)$ . For each  $G_j$ ,  $j = 1, 2$ , we simulate  $\theta_1, \dots, \theta_{500} \stackrel{iid}{\sim} G_j$  and  $\theta_i = 0$  for  $i = 501, \dots, 3000$  i.e.  $w = 1/6$ . Finally, for each  $G_j$  we simulate (i)  $y_i \stackrel{ind.}{\sim} N(\theta_i, 0.5^2)$  for the model with a common, unknown  $\sigma^2$ , or (ii)  $y_i \stackrel{ind.}{\sim} N(\theta_i, \sigma_i^2)$  where  $\sigma_i \sim \Gamma(5, 10)$  where  $E(\sigma_i) = 0.5$ . An equal-tailed 95% probability interval for  $\sigma_i$  is  $(0.16, 1.02)$ .

We continue by exploring the FDR and FNR of each scenario, across varying values of  $k$ ,  $J$ , and prior information on  $w$  as well as the estimates of the non-null densities. Then, we increase the sample size to  $n = 30,000$  to display the scalability of the model.

### 6.1 Simulation with $G_1$

We simulated 100 datasets for  $G_1$ , as described above, and applied our method with  $a_w = 1$ ,  $b_w = 1$ ,  $J = 5$ , and  $c$  random. As seen in Figure 2 our density estimation is accurate for both (i) unknown, common and (ii) known variances, particularly considering that only 500 of the 3000 data points come from the non-null distribution. This results in the method's ability to keep both FDR and FNR relatively low, as seen in the summary of FDP and FNP in Table 3.

Approximating Scott and Berger (2006) by setting  $J = 8$  and fixing  $c = 100,000$ , as reported in Table 4 we see that the results are similar, but the parametric approach has slightly smaller FDR.

|          | Known Variance |       |       | Common Variance |       |       |
|----------|----------------|-------|-------|-----------------|-------|-------|
| Cost $k$ | MNR            | FDR   | FNR   | MNR             | FDR   | FNR   |
| 1/10     | 248            | 0.025 | 0.093 | 234             | 0.024 | 0.098 |
| 1/3      | 264            | 0.039 | 0.090 | 250             | 0.038 | 0.094 |
| 1        | 315            | 0.106 | 0.081 | 301             | 0.100 | 0.084 |
| 3        | 422            | 0.247 | 0.070 | 399             | 0.227 | 0.074 |
| 10       | 518            | 0.347 | 0.065 | 481             | 0.320 | 0.069 |

Table 3: Errors Summary over 100 Simulations for  $G_1$ ,  $J = 5$  and  $c$  random under  $T_{i1}$ ; MNR is the mean number of rejections.

### 6.2 Simulation with $G_2$

We simulated 100 datasets for  $G_2$ , as described above, and applied our method. As we see in Figure 2 good density estimation is still enjoyed for this median-zero mixture of two normals. In Figure 2 the improvement of the Polytrees Nonparametric approach is clear; the density estimates of Scott and Berger (2006) miss the bimodal density and fit one gaussian density over it. The

|          | Known Variance |       |       | Common Variance |       |       |
|----------|----------------|-------|-------|-----------------|-------|-------|
| Cost $k$ | MNR            | FDR   | FNR   | MNR             | FDR   | FNR   |
| 1/10     | 232            | 0.013 | 0.098 | 219             | 0.016 | 0.102 |
| 1/3      | 243            | 0.021 | 0.095 | 233             | 0.023 | 0.098 |
| 1        | 279            | 0.058 | 0.087 | 272             | 0.061 | 0.089 |
| 3        | 344            | 0.146 | 0.077 | 339             | 0.150 | 0.079 |
| 10       | 514            | 0.342 | 0.065 | 504             | 0.340 | 0.067 |

Table 4: Errors Summary over 100 Simulations for  $G_1$ ,  $J = 8$  and  $c$  fixed at 100,000 under  $T_{i1}$ ; MNR is the mean number of rejections.

summary of FDP and FNP in Table 5 shows the ability to control FDR and FNR is preserved as well.

|          | Known Variance |       |       | Common Variance |       |       |
|----------|----------------|-------|-------|-----------------|-------|-------|
| Cost $k$ | MNR            | FDR   | FNR   | MNR             | FDR   | FNR   |
| 1/10     | 201            | 0.023 | 0.108 | 184             | 0.024 | 0.113 |
| 1/3      | 214            | 0.038 | 0.105 | 197             | 0.038 | 0.111 |
| 1        | 256            | 0.100 | 0.098 | 242             | 0.109 | 0.103 |
| 3        | 352            | 0.247 | 0.088 | 350             | 0.267 | 0.092 |
| 10       | 443            | 0.350 | 0.083 | 448             | 0.373 | 0.086 |

Table 5: Errors Summary over 100 Simulations for  $G_2$ ,  $J = 5$  and  $c$  random under  $T_{i1}$ ; MNR is the mean number of rejections.

|          | Known Variance |       |       | Common Variance |       |       |
|----------|----------------|-------|-------|-----------------|-------|-------|
| Cost $k$ | MNR            | FDR   | FNR   | MNR             | FDR   | FNR   |
| 1/10     | 187            | 0.015 | 0.112 | 176             | 0.016 | 0.115 |
| 1/3      | 196            | 0.022 | 0.110 | 185             | 0.023 | 0.113 |
| 1        | 225            | 0.056 | 0.103 | 212             | 0.061 | 0.108 |
| 3        | 278            | 0.145 | 0.096 | 263             | 0.149 | 0.101 |
| 10       | 420            | 0.337 | 0.086 | 399             | 0.334 | 0.090 |

Table 6: Errors Summary over 100 Simulations for  $G_2$ ,  $J = 8$  and  $c$  fixed at 100,000 under  $T_{i1}$ ; MNR is the mean number of rejections.

### 6.3 Simulation with $G_1$ Scalability Results

We simulated 100 datasets for  $G_1$ , this time with  $\theta_1, \dots, \theta_{5000} \stackrel{iid}{\sim} G_1$  and  $\theta_i = 0$  for  $i = 5001, \dots, 30000$ . Finally, we simulate (i)  $y_i \stackrel{ind.}{\sim} N(\theta_i, 0.5^2)$  for the model with a common, unknown  $\sigma^2$ , or (ii)  $y_i \stackrel{ind.}{\sim} N(\theta_i, \sigma_i^2)$  where  $\sigma_i \sim \Gamma(5, 10)$  where  $E(\sigma_i) = 0.5$ . Our density estimation is very good for both common and known variance; with the expanded data set we see tighter density estimates. This result shows the scalability of our method as the ability to keep both FDR and FNR relatively low, as seen in the summary of FDP and FNP in Table 7, is preserved.

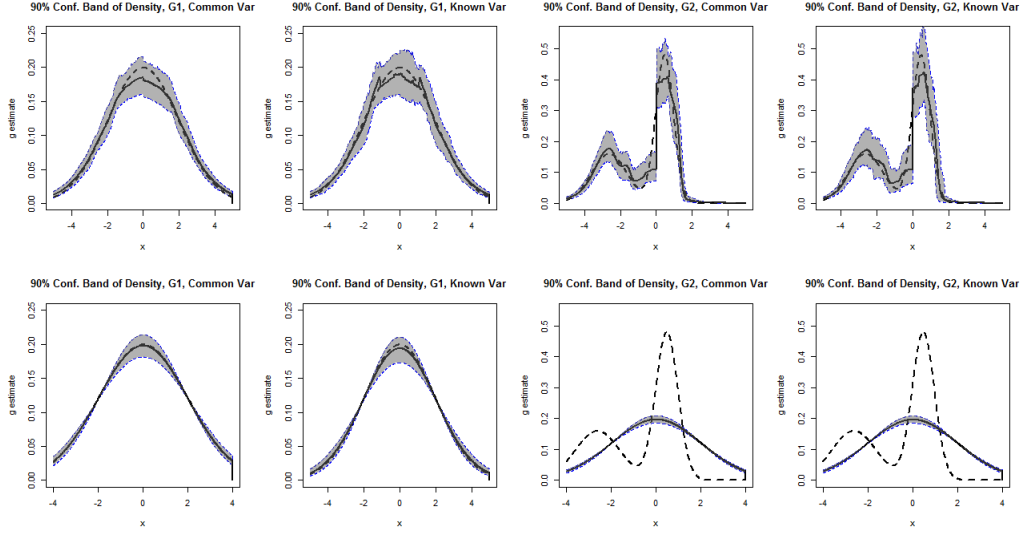


Figure 2: The top row shows simulation results for G1 and G2 using the nonparametric approach and the bottom row shows Scott and Berger (2006) results for the same data. The dotted densities are the true densities of the non-zero means and the solid density with the gray band is the estimated density with the 90 percent confidence band.

|          | Known Variance |       |       | Common Variance |       |       |
|----------|----------------|-------|-------|-----------------|-------|-------|
| Cost $k$ | MNR            | FDR   | FNR   | MNR             | FDR   | FNR   |
| 1/10     | 2499           | 0.025 | 0.093 | 2354            | 0.024 | 0.098 |
| 1/3      | 2659           | 0.040 | 0.090 | 2516            | 0.039 | 0.094 |
| 1        | 3177           | 0.107 | 0.081 | 3021            | 0.102 | 0.085 |
| 3        | 4272           | 0.253 | 0.070 | 4000            | 0.234 | 0.075 |
| 10       | 5229           | 0.352 | 0.065 | 4799            | 0.323 | 0.070 |

Table 7: Errors Summary over 100 Simulations for  $G_1$ ,  $J = 5$  and  $c$  random under  $T_{i1}$ ; MNR is the mean number of rejections.

## 6.4 Simulation with $G_2$ Scalability Results

We simulated 100 datasets for  $G_2$ , this time with  $\theta_1, \dots, \theta_{5000} \stackrel{iid}{\sim} G_2$  and  $\theta_i = 0$  for  $i = 5001, \dots, 30000$ . Finally, we simulate (i)  $y_i \stackrel{ind.}{\sim} N(\theta_i, 0.5^2)$  for the model with a common, unknown  $\sigma^2$ , or (ii)  $y_i \stackrel{ind.}{\sim} N(\theta_i, \sigma_i^2)$  where  $\sigma_i \sim \Gamma(5, 10)$  where  $E(\sigma_i) = 0.5$ . With the expanded data, tighter density estimation is still enjoyed for this median-zero mixture of two normals. As seen in the summary of FDP and FNP in Table 8 the ability to control FDR and FNR is preserved too.

## 6.5 Simulation with $G_1$ Varying Levels $w$

We simulated 100 datasets for  $G_2$ , with  $\theta_1, \dots, \theta_{500} \stackrel{iid}{\sim} G_2$  and  $\theta_i = 0$  for  $i = 501, \dots, 3000$ . Finally, we simulate  $y_i \stackrel{ind.}{\sim} N(\theta_i, \sigma_i^2)$  where  $\sigma_i \sim \Gamma(5, 10)$  where  $E(\sigma_i) = 0.5$ . This time, we gave prior information  $\hat{p}$ , on  $w$ .

As seen in Table 9, when we give perfect information  $\hat{p} = 0.5$  on  $w$  we see very similar results to our default methodology with no prior information on  $w$ . As we decrease  $\hat{p}$ , we see that fewer

|          | Known Variance |       |       | Common Variance |       |       |
|----------|----------------|-------|-------|-----------------|-------|-------|
| Cost $k$ | MNR            | FDR   | FNR   | MNR             | FDR   | FNR   |
| 1/10     | 1837           | 0.018 | 0.113 | 2031            | 0.021 | 0.108 |
| 1/3      | 1951           | 0.031 | 0.111 | 2151            | 0.034 | 0.105 |
| 1        | 2356           | 0.092 | 0.103 | 2561            | 0.096 | 0.098 |
| 3        | 3293           | 0.236 | 0.093 | 3526            | 0.244 | 0.088 |
| 10       | 4141           | 0.335 | 0.087 | 4462            | 0.351 | 0.082 |

Table 8: Errors Summary over 100 Simulations for  $G_2$ ,  $J = 5$  and  $c$  random under  $T_{i1}$ ; MNR is the mean number of rejections.

observations are rejected which we expect as our prior information on  $w$  intimates that we should be rejecting less. The most telling piece of this simulation are the density estimates seen in Figure 3; as the prior on  $w$  decreases, we see the mass around zero is slowly removed finally yielding “interestingly different” rejected observations and a density estimate of those observations as seen in 3.

| $\hat{p}$ Value | MNR | FDR   | FNR   |
|-----------------|-----|-------|-------|
| 500/3000        | 316 | 0.107 | 0.081 |
| 315/3000        | 293 | 0.077 | 0.084 |
| 248/3000        | 284 | 0.066 | 0.086 |
| 100/3000        | 260 | 0.039 | 0.091 |

Table 9: Errors Summary over 100 Simulations for  $G_2$  with known variance,  $m = 10,000$ ,  $k = 1$ ,  $J = 5$  and  $c$  random under  $T_{i1}$ ; MNR is the mean number of rejections.

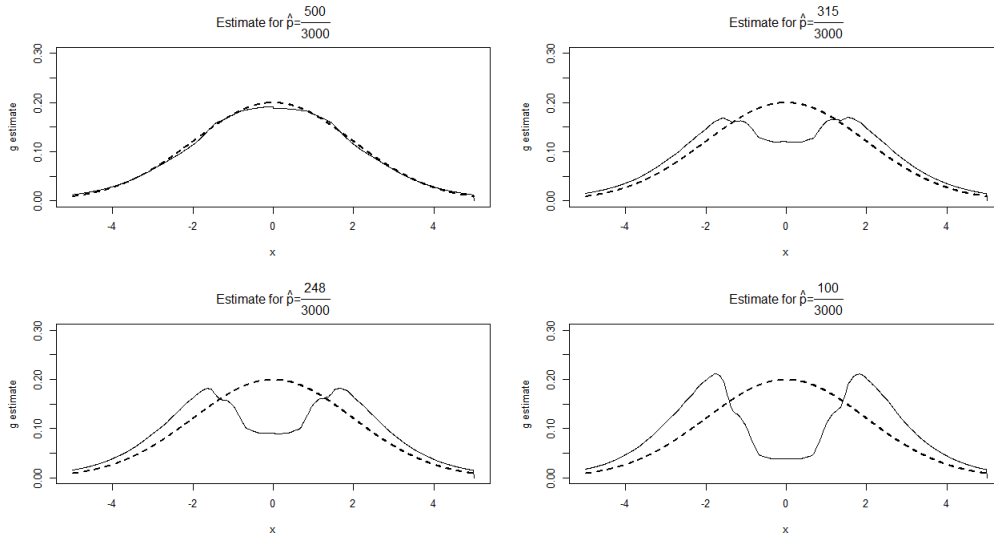


Figure 3: Density Estimates over different values of  $\hat{p}$  are the solid density estimates and density estimate with no prior on  $w$  is the dotted estimate.

## 6.6 Simulation with $G_2$ Varying Levels $J$

We simulated 100 datasets for  $G_2$ , with  $\theta_1, \dots, \theta_{500} \stackrel{iid}{\sim} G_2$  and  $\theta_i = 0$  for  $i = 501, \dots, 3000$ . Finally, we simulate  $y_i \stackrel{ind.}{\sim} N(\theta_i, \sigma_i^2)$  where  $\sigma_i \sim \Gamma(5, 10)$  where  $E(\sigma_i) = 0.5$ . This time, we varied the number of levels,  $J$ , in the Polya tress.

As seen in Table 10, we see that FDR and FNR vary little over different levels of  $J$ . This shows robustness to the choice  $J$ ; i.e. a “leveling off” noted by Hanson (2006).

| Levels $J$ Value | MNR | FDR   | FNR   |
|------------------|-----|-------|-------|
| $J = 5$          | 242 | 0.109 | 0.103 |
| $J = 6$          | 256 | 0.101 | 0.098 |
| $J = 7$          | 256 | 0.101 | 0.098 |
| $J = 8$          | 256 | 0.101 | 0.098 |

Table 10: Errors Summary over 100 Simulations for  $G_2$  with Known Variance,  $k = 1$  and  $c$  random under  $T_{i1}$ ; MNR is the mean number of rejections.

## 7 Data Analyses

Below we apply our approach to three different hypothesis testing scenarios: a difference of proportions, a paired difference, and a two sample difference.

### 7.1 Proportional Difference

Sun and McLain (2012) provide an example using educational survey data from the Adequate Yearly Progress study on the academic performances of students across different social and fiscal demographics; we are interested in schools where the difference is interesting.

In the case of the usual Z-statistic for proportions we take  $y_i = \hat{p}_{SA_i} - \hat{p}_{SD_i}$  where  $\hat{p}_{SA_i}$  is the proportion of social-economically advantaged students at school  $i$  and  $\hat{p}_{SD_i}$  is the proportion of social-economically disadvantaged students at school  $i$  with  $\sigma_i^2 = \hat{p}_{SA_i}(1 - \hat{p}_{SA_i})/n_{SA_i} + \hat{p}_{SD_i}(1 - \hat{p}_{SD_i})/n_{SD_i}$  for  $i = 1, \dots, 7866$ .

In this particular case, the students do differently on average so we have to shift our observations to be mean zero. Here, we take shifted  $y_{shifted_i} = \hat{p}_{SA_i} - \hat{p}_{SD_i} - \eta$  where  $\eta$  is the median of all  $y_i$ ’s. The usual T-statistic approach yields 5,731 and 5,246 rejections at the 0.10 and 0.05 level of significance respectively. Our default, no prior information on  $w$ , approach estimates  $w = .9558$  with a 95% credible interval of the number of non-null observations to be (6801, 7851).

In Table 11 we see the results of our model without using any prior information on  $w$ . As we decrease our cost  $k$  we see that we reject the null of fewer observations by design of threshold  $T_{i2}$  and with cost  $k = 1$  we see a similar number of rejections to the usual approach at the 0.10 and .05 levels. The non-null density estimate, in this case, is approximately bell-shaped and symmetric at zero.

| Cost                 | 1/10 | 1/3  | 1    | 3    | 10   |
|----------------------|------|------|------|------|------|
| Number of Rejections | 1863 | 2298 | 4468 | 7511 | 7736 |

Table 11: Rejection Summary over various costs

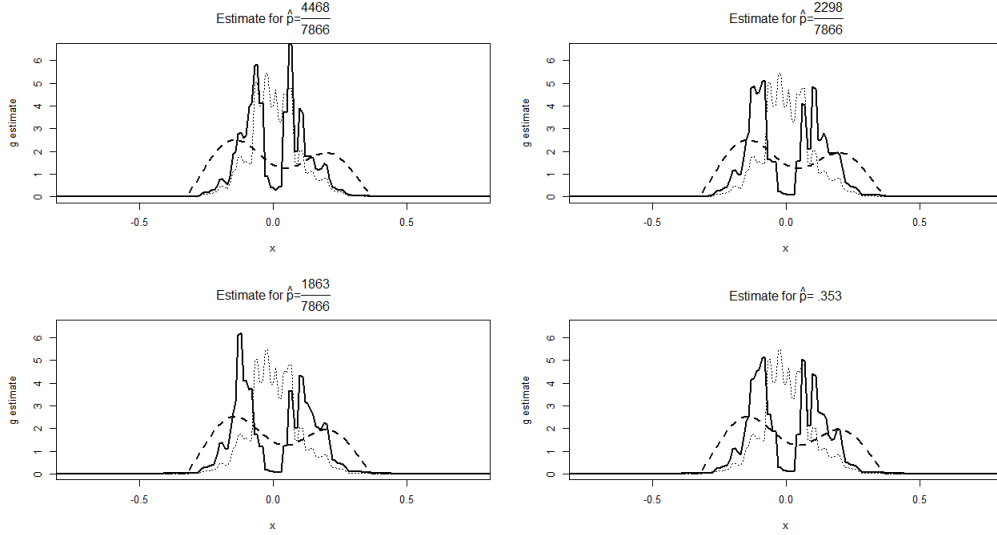


Figure 4: Density Estimates over different values of  $\hat{p}$  are the solid density estimates and density estimate with no prior on  $w$  is the dotted estimate. The density estimate from Sun and McLain (2012) is represented by the dashed density estimate.

Sun and McLain (2012) explore this data, searching for schools that are “interestingly different.” We can consider this problem by inputting prior information on  $w$ . Sun and McLain (2012) consider the “oracle” method in Jin (2008) by using the  $\hat{w} = .353$  which we can use as the prior information on  $w$ . Using this prior information we report a 95% credible interval of the number of non-null observations to be (3366, 3545) and the non-null density estimates, seen in Figure 4, are bimodal indicating that there are two groups of “interestingly” different schools.

In the absence of prior information, or the use of an “oracle” method, similar results could be achieved by using the proportion of rejections in our default, no prior information on  $w$ , approach for small values of  $k$ . As seen in Figure 4, using the proportions created by the number of rejections at  $k = 1$ ,  $k = 1/3$  and  $k = 1/10$  we obtain results similar to Sun and McLain (2012).

## 7.2 Paired Difference

Next, we consider the gene expression data from the microarray experiments of Colon tissue samples of Notterman et al. (2001). This data, available through the Princeton University Gene Expression Project, consists of 7,457 genes measurements for 18 patients on both tumor and normal tissues.

In the case of the usual, paired Student T statistic we take  $y_i = \bar{x}_{d_i}$  where  $\bar{x}_{d_i}$  is the mean pairwise difference and  $\sigma_i^2 = s_{d_i}^2/18$  for  $i = 1, \dots, 7457$ .

Again, we take shifted  $y_{shifted_i} = \bar{x}_{d_i} - \eta$  where  $\eta$  is the median of all  $y_i$ ’s. The usual T-statistic approach yields 2,818 and 2,205 rejections at the 0.10 and 0.05 level of significance respectively. Our default, no prior information on  $w$ , approach estimates  $w = .3621$  with a 95% credible interval of the number of non-null observations to be (2577, 2839).

In Table 12 we see the results of our model without using any prior information on  $w$ . As we decrease our cost  $k$  we see that we reject the null of fewer observations by design of threshold  $T_{i1}$ . The non-null density estimate, in this case, is approximately bell-shaped and symmetric at zero.

|                      |      |     |      |      |      |
|----------------------|------|-----|------|------|------|
| Cost                 | 1/10 | 1/3 | 1    | 3    | 10   |
| Number of Rejections | 740  | 824 | 1029 | 1346 | 1533 |

Table 12: Rejection Summary over various costs

### 7.3 Two Sample Difference

Ausin et al. (2011) consider the gene expression data from the microarray experiments of Colon tissue samples of Alon et al. (1999). This data, available through the R package “plsge-nomics” consists of 2,000 genes for 62 samples; 40 of these samples are from tumor tissues and 22 are from normal tissues. It should be noted that the original dataset consisted of 6,500 genes and the 2,000 observations in the available data are the genes with highest minimal intensity.

In the case of the usual, two sample Student T statistic we take  $y_i = \bar{x}_{norm_i} - \bar{x}_{tumor_i}$  and  $\sigma_i^2 = s_{norm_i}^2/22 + s_{tumor_i}^2/40$  for  $i = 1, \dots, 2,000$ .

Again, we take shifted  $y_{shifted_i} = \bar{x}_{norm_i} - \bar{x}_{tumor_i} - \eta$  where  $\eta$  is the median of all  $y_i$ ’s. The usual T-statistic approach yields 626 and 477 rejections at the 0.10 and 0.05 level of significance respectively. Ausin et al. (2011) report 223 differentially expressed genes with posterior probability one. Our default, no prior information on  $w$ , approach estimates  $w = .7178$  with a 95% credible interval of (1139, 1801).

In Table 13 we see the results of our model without using any prior information on  $w$ .

|                      |      |     |     |     |     |
|----------------------|------|-----|-----|-----|-----|
| Cost                 | 1/10 | 1/3 | 1   | 3   | 10  |
| Number of Rejections | 126  | 159 | 240 | 467 | 633 |

Table 13: Rejection Summary over various costs

## 8 Java Applet

An increasing trend among statisticians is to provide R packages for fitting complex methodology. These packages allow others familiar with the R computing environment to implement methods otherwise not readily available, and hence allow the routine use of new methodology. However, many scientists in other fields are unfamiliar with R and other programming languages. These scientists often use specialized software or Excel to implement statistical methods.

Perhaps the most widely available venue for the dissemination of new statistical methods is through the use of online Java applets that are compatible with any machine that can run Java; it is in this spirit that we provide such a Java applet. Scientists of all backgrounds are able to utilize this methodology without the steep learning curve of learning a new programming language. Figure 5 shows a screenshot of the Java applet that we created.

### 8.1 Input

The left side of the applet has text areas for input. The top text area is for entering the observations; each observation should have its own row. The applet is smart enough to know whether or not the input contains known variances and in the given variance case we expect the observation and variance to be comma separated; again, each couple should have its own row. In the second row, the two text areas are for the desired the number of Gibbs iterates and the number of burn in iterates, respectively. In the third row, the two text areas are for the desired cost  $k$  and number of levels for the Polya tree  $J$ . The following line for input asks whether or not prior

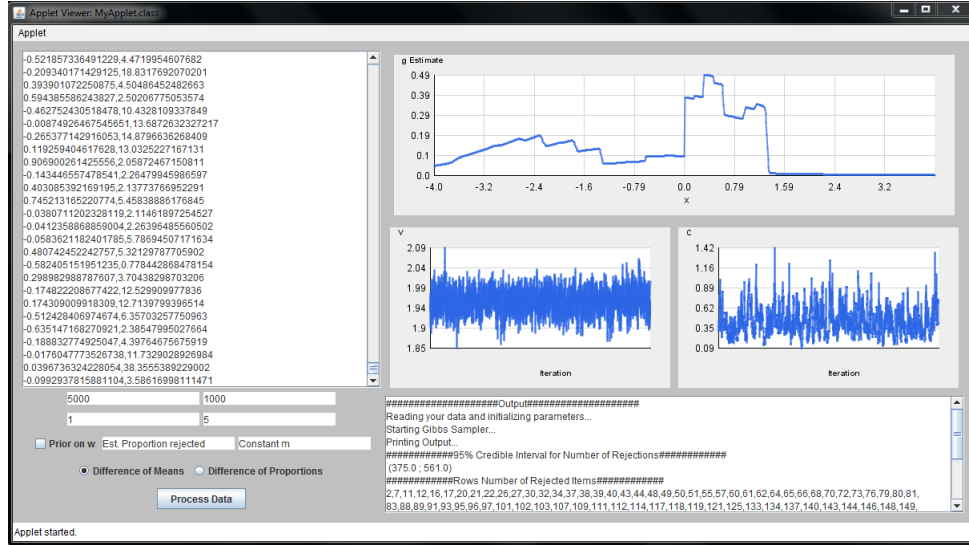


Figure 5: Screenshot of the Java Application with Results

information on  $w$  will be used followed by two text areas - the estimated proportion of rejections and the constant  $m$ . The last line inquires whether the test is a difference of means or differences of proportions so that the program can toggle between thresholds  $T_{i1}$  and  $T_{i2}$ .

## 8.2 Ouput

The right side of the applet displays output. The first graph is the density estimate for the non-zero means, and the next two show the variables  $v$  and  $c$  after burn in. If the input is set correctly these graphs should appear fairly consistent over the iterates. If the graphs are not consistent, burn in and Gibbs iterates may need to be increased. The textual output gives a credible interval for the number of observations rejected and a list of which observations were rejected overall according to the Threshold.

This applet is available through <http://www.stat.sc.edu/~cipolli/BMT/BMT.html>.

## 9 Conclusion

The suggested approximate finite Polya tree multiple testing procedure is very successful in correctly classifying the observations with non zero mean. This holds even when the non zero means are simulated from a mean zero distribution, as seen in our simulation of  $\theta_i$  from a  $N(0, 2^2)$ , which is particularly impressive as we can expect many of these ‘non-zero’ means to be very close to zero.

The performance of the multiple comparisons was evaluated using FDR and FNR, both of which were kept low during simulation for relatively small and large numbers of observations. A nice aspect of the methodology and Java applet is that we are able to provide an approximation of the density of the non zero means that is very close to the actual density function. Further we are able to do this for “interestingly different” observations for the cases where that is of interest.

This model assumes that  $\theta_i$  and  $\sigma_i^2$  are independent and is sensitive to the prior specifications including  $w$ . Our model is also sensitive to the data being median zero; simulation results, not included, show that deviations from this lead to inflated error.

## References

- (2008). Technical trading revisited: False discoveries, persistence tests, and transaction costs.
- Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natn. Acad. Sci. USA*, 96(12):6745–6750.
- Ausin, M., Gomez-Villegas, M., Gonzalez-Perez, B., Rodriguez-Bernal, M., Salazer, I., and Sanz, L. (2011). Bayesian analysis of multiple hypothesis testing with applications to microarray experiments. *Communications in Statistics - Theory and Methods*, 40(13):2276–2291.
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, 57(1):289–300.
- Blomquist, J. (2014). Multiple inference and market integration: An application to Swedish fish markets. *Journal of Agricultural Economics*, 66(1):221–234.
- Chumbley, J. R. and Friston, K. J. (2009). False discovery rate revisited: FDR and topological inference using Gaussian random fields. *NeuroImage*, 44(11):62–70.
- Do, K., Müller, P., and Tang, F. (2005). A Bayesian mixture model for differential gene expression. *Journal of the Royal Statistical Society*, 54(03):627–644.
- Efron, B., Tibshirani, R., Storey, J., and Tusher, V. (2001). Empirical bayes analysis of microarray experiment. *Journal of the American Statistical Association*, 96(456):1151–1160.
- Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics*, 02(04):615–629.
- Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery procedure. *Journal of the Royal Statistical Society. Series B*, 64(3):499–517.
- Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242.
- Hanson, T. (2006). Inference for mixtures of finite Polya tree models. *Journal of the American Statistical Association*, 101(0476):1548–1565.
- Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance comparing individual means in the analysis of variance. *Biometrika*, 75(4):800–802.
- Jin, J. (2008). Proportion of non-zero normal means: universal oracle equivalences and uniformly consistent estimators. *Journal of the Royal Statistical Society*, 70(3):461–493.
- Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modelling. *The Annals of Statistics*, 20(03):1222–1235.
- Lavine, M. (1994). More aspects of Polya tree distributions for statistical modelling. *The Annals of Statistics*, 22(03):1161–1176.
- Mauldin, R. D., Sudderth, W. D., and Williams, S. C. (1992). Polya trees and random distributions. *The Annals of Statistics*, 20(03):1203–1221.

- Müller, P., Parmiani, G., and Rice, K. (2007). FDR and Bayesian multiple comparisons rules. *Bayesian Statistics*, 8:349–370.
- Notterman, D. A., Alon, U., and Sierk, A. J. (2001). Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. *Cancer Research*, 61(7):3124–3130.
- Peña, E. A., Habiger, J. D., and Wu, W. (2011). Power-enhanced multiple decision functions controlling family-wise error and false discovery rates. *The Annals of Statistics*, 39(1):556–583.
- Scott, J. G. and Berger, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, 136(07):2144–2162.
- Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633.
- Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B*, 64(3):479–498.
- Sun, W. and McLain, A. C. (2012). Multiple testing of composite null hypotheses in heteroscedastic models. *Journal of the American Statistical Association*, 107(398):673–687.
- Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, 5(2):99–114.
- Wu, W. and Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, 102(479):901–912.