

# FALSE DISCOVERY RATE AND EMPIRICAL NULL METHODS

OMKAR MURALIDHARAN

## 1. INTRODUCTION

A common problem is to find the the most interesting items in a large pool of prospects for investigation. A biologist might search for differentially expressed genes; a trader might look for lucrative trading positions. One way to tackle this problem is by classifying each prospect as boring (null) or interesting (non-null), that is, conduct hypothesis tests. The sheer number of hypotheses, however, creates new problems. This essay will look at one multiple hypothesis testing technique for this situation, false discovery rates.

There are many ways to classify prospects. The first step in choosing one is to pick the relevant measure of error. One choice is the probability of making at least one type I error. This measure, called the family-wise error rate is most appropriate when the tests will be used together in such a way that even one false declaration of significance can be very costly. The exploratory nature of our problem suggests that the FWER is too strict - when we look for interesting prospects, we usually plan to investigate them further. This means that we might be willing to accept a few more false positives in order to get more true positives. The expected proportion of false positives, called the false discovery rate (FDR), is thus a more natural choice for this kind of problem. If the null distribution of the scores is known, there are simple procedures to control the FDR, and these procedures are in some sense optimal.

Often, however, the null distribution of the scores is unknown, or the theoretical null distribution is unrealistic. In this situation, it makes sense to estimate the null distribution from the data, giving an “empirical null.” Doing this can be very useful, but it also raises new questions about how to evaluate an empirical null method, and how to find the best one for any situation.

In this paper, I will review the main points of standard false discovery theory, which assumes the null distribution is known. I will then look at the first empirical null methods, and introduce a new method based on symmetry. After looking at the power and FDR control of these methods in a few examples, I will see how they perform when the true null distribution is gradually skewed. Next, I will see how well the methods estimate the rejection region we would use if we knew the true null, and will argue that this is a better measure of error than power or control. Finally, I will consider a Bayesian approach to estimating the null in the hope that this might point toward an optimal empirical null method.

## 2. BASIC FDR RESULTS

In this section, I will review some basic results that are useful in false discovery work. Benjamini and Hochberg gave the first practical procedure to control the FDR in [6]. First, however, we need to state the problem more precisely. We want to simultaneously test  $m$  null hypotheses  $H_1, \dots, H_m$  at level  $\alpha$ , giving p-values  $p_i$ ,

and  $m_0$  of these hypotheses are truly null. Typically, each hypothesis corresponds to some sort of test object, like a gene, and the null hypothesis is that the object is not interesting enough for further study. This means that null status may not be random, but we will often think of null status as being random. This makes little difference for large  $m$ .

For any testing procedure based on the p-values, we can consider the rejection region  $\Gamma$ . Let  $R$  be the number of rejections, and  $V$  the number of rejections of truly null hypotheses. Intuitively, we want to define the FDR by  $FDR = \frac{V}{R}$ , but  $R$  may be 0. Also, the true values of  $V$  and  $R$  are unknown, and we cannot hope to control  $FDR$  if it is directly defined in terms of these two quantities: if  $m_0 = m$  and  $R > 0$ , the  $FDR$  will be 1, so the only way to control the  $FDR$  would be to never reject any hypotheses. To fix this, Benjamini and Hochberg define  $FDR$  to be 0 when  $R = 0$  and focus on its expectation:  $FDR = E(\frac{V}{R \vee 1}) = E(\frac{V}{R} | R > 0) P(R > 0)$ . The  $FDR$  is a function of the rejection region  $\Gamma$ , although the region will often be clear from context. Benjamini and Hochberg then give the following procedure for controlling the  $FDR$ .

**Theorem 2.1** (Benjamini and Hochberg). *Suppose we test  $m$  hypotheses based on the ordered p-values  $p_{(1)}, \dots, p_{(m)}$ . Let  $H_{(i)}$  be the hypothesis corresponding to each  $p_{(i)}$ . For a given  $q$ , let  $k = \max\{i : p_{(i)} \leq \frac{i}{m}q\}$ . Let the rejection region of the test be  $\Gamma = [0, p_{(k)}]$ , so  $H_{(1)}, \dots, H_{(k)}$  are rejected. Then  $E(FDR(\Gamma)) \leq q$ .*

In [8], Storey points out that this approach fixes a desired false discovery rate, then finds a rejection region that has a smaller false discovery rate on average. He points out a few weaknesses in this approach. First, no information about  $m_0$  is used. This is wasteful, since the p-values contain information about  $m_0$ . Second, and more seriously, treating the problem in this way obscures the behavior of the realized false discovery rate. Although the procedure guarantees  $E(FDR) \leq q$  for the generated rejection region, the region is a random quantity, so the actual  $FDR$  may vary. Instead, Storey treats the problem as an estimation problem - first fix a rejection region and then estimate the  $FDR$  for that region. This allows us to look at many different rejection regions, and to use the bootstrap to generate confidence intervals for  $FDR$ .

Storey also argues in [9] that we should consider  $E(\frac{V}{R} | R > 0)$  instead of the  $FDR$  - he calls this the  $pFDR$ , or positive false discovery rate. He believes this is a better measure of error, since it is equal to 1 when there are no false hypotheses ( $m_0 = m$ ), and also since we are usually not interested in situations where  $R = 0$ . The  $pFDR$  also has an attractive Bayesian interpretation, which is given below. The distinction between  $FDR$  and  $pFDR$  is relatively minor, since  $m$  is usually large enough that  $P(R > 0)$  is close to 1 for most rejection regions.

**Theorem 2.2** (Storey). *Suppose we have  $m$  hypotheses  $H_1, \dots, H_m$  with corresponding statistics  $T_1, \dots, T_m$ , and we fix a rejection region  $\Gamma$ . Let  $H_i$  be 0 or 1 if the null hypothesis is true or false. Suppose that  $(T_i, H_i)$  are iid, with  $P(H_i = 0) = \pi_0$  and  $T_i | H_i \sim (1 - H_i) F_0 + H_i F_1$ , where  $\pi_0$  is the proportion of null hypotheses, and  $F_i$  are the null and alternative distributions. Then  $pFDR(\Gamma) = P(H_i = 0 | T_i \in \Gamma)$ .*

*Proof.* We have

$$\begin{aligned} pFDR(\Gamma) &= E\left(\frac{V}{R} \mid R > 0\right) \\ &= \sum_{i=1}^m E\left(\frac{V}{i} \mid R = i\right) P(R = i \mid R > 0). \end{aligned}$$

Also, we have

$$E(V \mid R = i) = E\left(\sum_{j=1}^m 1(T_j \in \Gamma) 1(H_j = 0) \mid R = i\right).$$

Since the statistics are iid, we can simplify this to

$$\begin{aligned} E(V \mid R = i) &= E\left(\sum_{j=1}^m 1(T_j \in \Gamma) 1(H_j = 0) \mid T_1, \dots, T_i \in \Gamma, T_{i+1}, \dots, T_m \notin \Gamma\right) \\ &= E\left(\sum_{j=1}^i 1(H_j = 0) \mid T_1, \dots, T_i \in \Gamma, T_{i+1}, \dots, T_m \notin \Gamma\right) \\ &= \sum_{j=1}^i E(1(H_j = 0) \mid T_j \in \Gamma) \\ &= i P(H_i = 0 \mid T_i \in \Gamma). \end{aligned}$$

Putting this back into the original expression gives

$$\begin{aligned} pFDR(\Gamma) &= P(H_i = 0 \mid T_i \in \Gamma) \sum_{i=1}^m P(R = i \mid R > 0) \\ &= P(H_i = 0 \mid T_i \in \Gamma) \end{aligned}$$

as desired.  $\square$

Storey notes that the theorem approximately holds for nonrandom  $H_i$  if  $m$  is large, and that we can consider composite alternatives by taking  $F_1$  to be the mixture distribution of the alternatives.

This result of the theorem can be restated slightly: under the conditions above,

$$pFDR(\Gamma) = \pi_0 \frac{P(T_i \in \Gamma \mid H_i = 0)}{P(T_i \in \Gamma)}.$$

Under the null hypothesis,  $T_i$  has known distribution, so this suggests the estimate

$$p\hat{FDR} = \pi_0 \frac{P(T_i \in \Gamma \mid H_i = 0)}{\frac{\#\{T_i \in \Gamma\}}{m}}.$$

Typically  $\pi_0$  is also unknown, but we can estimate it. Before giving this estimate, it is convenient to assume that we are working with p-values -  $T_i \in [0,1]$ , given  $H_i = 0$ ,  $T_i \sim \text{Unif}(0,1)$ , and  $\Gamma$  is of the form  $[0, \gamma]$  (since p-values are constructed in terms of nested rejection regions, it never makes sense to reject for a value and accept for a smaller one).

To estimate  $\pi_0$ , we need to make a further assumption in order to distinguish null and non-null values. We assume that our test has reasonable power against the non-nulls, so that most of the values near 1 come from null cases. With this assumption, we can use the values near 1 and our knowledge that the null distribution is  $\text{Unif}(0,1)$

to estimate the proportion of null cases. Assume that all values right of  $\lambda$  are from null hypotheses. We expect  $\pi_0 m$  null cases, so there should be  $\pi_0 m(1 - \lambda)$  cases right of  $\lambda$ . Rearranging this gives us the estimate

$$\hat{\pi}_0 = \frac{\#\{T_i > \lambda\}}{(1 - \lambda)m}.$$

For ease of notation I will suppress the dependence of  $\hat{\pi}_0$  on  $\lambda$ . The choice of  $\lambda$  is a bias-variance tradeoff: large  $\lambda$  will give lower bias, and smaller  $\lambda$  will give lower variance. The bootstrap can be used to pick an optimal  $\lambda$ , as we will see later.

Plugging this estimate into our estimate of  $p\hat{FDR}$  gives

$$p\hat{FDR} = \frac{\#\{T_i > \lambda\}}{(1 - \lambda) \#\{T_i \in \Gamma\}} P(T_i \in \Gamma | H_i = 0).$$

Using our assumption that we are working with p-values, this simplifies to

$$p\hat{FDR} = \frac{\#\{T_i > \lambda\} \gamma}{(1 - \lambda) \#\{T_i < \gamma\}}.$$

For small  $\gamma$  however, there is a good chance that no  $T_i$  will be less than  $\gamma$ . We can just replace  $\#\{T_i < \gamma\}$  by 1 in this case. Doing this, however, brings us back to our old problem of how to define  $FDR$  when  $R = 0$ . Setting  $\#\{T_i < \gamma\} = 1$  when it is 0 amounts to using the  $FDR$  instead of the  $pFDR$ . This means we will underestimate  $pFDR$  by a factor of  $P(R > 0)$ . Storey notes that since we assume non-null cases are more likely to lead to rejections,  $P(R > 0)$  will be bounded below by  $P(T_i \geq \gamma | H_i = 0)^m = 1 - (1 - \gamma)^m$ , so we can divide by this bound. This gives the estimate

$$p\hat{FDR} = \frac{\#\{T_i > \lambda\} \gamma}{(1 - \lambda)(\#\{T_i < \gamma\} \vee 1)(1 - (1 - \gamma)^m)}.$$

Finally, since  $FDR$  is not conditioned on  $R > 0$ , we have the estimate

$$F\hat{DR} = \frac{\#\{T_i > \lambda\} \gamma}{(1 - \lambda)(\#\{T_i < \gamma\} \vee 1)}.$$

The two quantities are close for large  $m$ .

Now that we have these estimates, there are two questions we might ask. First, we might ask how good the estimators are. It turns out that they are quite good. Storey shows that the estimators have good finite sample properties [8], and also shows that they are closely related to a maximum likelihood estimator, and hence have asymptotically smallest variance for each level of bias (determined by choice of  $\lambda$ ). Both  $F\hat{DR}$  and  $p\hat{FDR}$  are modified versions of the estimator  $\frac{\hat{\pi}_0 \gamma}{\#\{T_i < \gamma\}}$ , and this is in fact a maximum likelihood estimator [8].

**Theorem 2.3.**  $\frac{\hat{\pi}_0 \gamma}{\#\{T_i < \gamma\}}$  is the maximum likelihood estimator of  $(1 + \frac{\pi_1(1 - F_1(\lambda))}{\pi_0(1 - \lambda)})pFDR(\gamma)$ , which equals  $(1 + \frac{P(T_i > \lambda | H_i = 1)}{P(T_i > \lambda | H_i = 0)})pFDR(\gamma)$ .

Second, we might ask if we can use our estimators to solve our original hypothesis testing problem. Intuitively, we want to look at  $F\hat{DR}$  for various  $\gamma$ , and pick the rejection threshold  $\gamma$  so that  $F\hat{DR}$  is below our desired cutoff. It is not clear that this procedure will control  $FDR$ . The following results proved by Storey, Taylor and Siegmund [7], however, show that this is in fact the case.

Before we see the results, let us first state the cutoff procedure precisely. Fix a desired  $FDR$  threshold  $q$ , and pick a value  $\lambda$  for the tuning parameter  $\lambda$ . We

need to make a small modification to  $F\hat{D}R$  if  $\lambda > 0$ . To ensure that  $\hat{\pi}_0 > 0$ , set  $\hat{\pi}_0 = \frac{\#\{T_i > \lambda\} + 1}{(1-\lambda)m}$  - this makes little difference for large  $m$ . Also, since we assumed all cases right of  $\lambda$  were null, set  $F\hat{D}R = 1$  for  $\gamma > \lambda$ . These changes are in fact not required for asymptotic  $FDR$  control, just for finite samples. Now our rejection procedure. We first find the cutoff point  $t$  by increasing  $\gamma$  until  $F\hat{D}R$  becomes too large:  $t = \max\{\gamma : F\hat{D}R \leq q\}$ . We then reject all hypotheses with p-values less than  $t$ . Let  $FDR^*$  be the  $FDR$  for the rejection region generated by this rule.

**Theorem 2.4.** *Suppose the  $T_i$  corresponding to  $H_i = 0$  are independent. Then  $E(FDR^*) \leq q$ .*

The result also holds under dependence, as long as the empirical cdfs of the null and alternative statistics converge pointwise almost surely to continuous distributions [7].

It may be surprising that a philosophically different approach based on estimation can also control the  $FDR$ . In fact, the next theorem, also from [7], shows a close connection between the Benjamini-Hochberg procedure and Storey's estimation approach.

**Theorem 2.5.** *Fix  $\lambda$ . Then the thresholding procedure is equivalent to the Benjamini-Hochberg procedure with  $m$  replaced by  $\hat{\pi}_0 m$ . In particular, for  $\lambda = 0$ , the two procedures are the same.*

*Proof.* Let  $k$  be the  $k$  from the Benjamini-Hochberg procedure with  $m$  replaced by  $\hat{\pi}_0 m$ , so  $k = \max\{i : T_{(i)} \leq \frac{i}{\hat{\pi}_0 m} q\}$ . Consider  $F\hat{D}R([0, T_{(i)}])$ . We know  $\#\{T_i \leq T_{(i)}\} = i$ , so  $F\hat{D}R([0, T_{(i)}]) = \frac{\hat{\pi}_0 m T_{(i)}}{i}$ . Hence  $k = \max\{i : F\hat{D}R(T_{(i)}) \leq q\}$ . this means the Benjamini-Hochberg procedure with  $m$  replaced by  $\hat{\pi}_0 m$  is the same as the thresholding procedure. For the second part, note that for  $\lambda = 0$ ,  $\hat{\pi}_0 = 1$ , so the replacement leaves the Benjamini-Hochberg procedure unchanged.  $\square$

Storey also gives a few more results of interest. He defines a quantity called the q-value, analogous to the p-value in hypothesis testing. The q-value of an observation is the smallest  $pFDR$  we must accept in order to reject that hypothesis. More formally, if  $A$  is a collection of nested rejection regions (for example,  $\{[0, \gamma] | \gamma \in [0, 1]\}$ ), the the q-value of an observation  $T_i$  with value  $t$  is  $\inf_{t \in \Gamma \in A} pFDR(\Gamma)$ . The q-value thus gives a measure of the false discovery error that comes from rejecting an observation with value  $t$ . Storey shows in [7] that the obvious estimate  $\inf_{t \in \Gamma \in A} pF\hat{D}R(\Gamma)$  is asymptotically simultaneously (in  $t$ ) conservative.

The q-value may lead us to consider another quantity, the more basic "local  $FDR$ ":  $fdr = P(H_i = 0 | T_i = t_i)$ . We can think of this quantity either as specific to the observations (looking at its values at the observed  $T_i$ ) or more generally like the q-values (looking at  $P(H_i = 0 | T_i = t)$  and using our iid assumption).

Another motivation for the local  $FDR$  comes from classification theory [9]. We can look at our multiple testing problem as a classification problem, where hypotheses need to be classified into null and non-null groups. Suppose we have no loss for a correct classification, a loss of  $w$  for a type II error, and a loss of  $1 - w$  for a type I error. Then the expected loss for any rejection region  $\Gamma$  is the Bayes error

$$BE = w P(H_i = 1, T_i \notin \Gamma) + (1 - w) P(H_i = 0, T \in \Gamma).$$

Now, analogous to the  $FDR$  and  $pFDR$ , define the missed discovery rate (expected proportion of false negatives) and the positive missed discovery rate (the

$MDR$  conditioned on at least one acceptance). Just as for the  $pFDR$ , we have the relationship  $pMDR = P(H_i = 1|T_i \notin \Gamma)$ . Then the Bayes error is

$$BE = w P(T_i \notin \Gamma) pMDR + (1 - w) P(T_i \in \Gamma) pFDR.$$

So minimizing the Bayes error is the same as minimizing a weighted sum of the  $pFDR$  and  $pMDR$ . It turns out, though, that it is enough to look at the local  $FDR$  to accomplish this minimization [9].

**Theorem 2.6.** *Assume the conditions of Theorem 2.2, and that  $F_0, F_1$  are continuous, with densities  $f_i$  and have common support. Then minimizing the Bayes error over arbitrary  $\Gamma$  or any weighted sum of  $pFDR$  and  $pMDR$  is equivalent to minimizing over the family*

$$\Gamma_k = \{t | fdr(t) \geq k\}.$$

*Proof.* The function  $x \mapsto \frac{x}{1-x}$  monotonically maps  $[0, 1]$  to  $[0, \infty]$ , so the family of  $\Gamma_k$  are equivalently given by  $\{t | \frac{\pi_1}{\pi_0} \frac{fdr(t)}{1-fdr(t)} \geq k\}$  for  $k \geq 0$ . We also have

$$\begin{aligned} fdr(t) &= P(H_i = 0 | T_i = t) \\ &= \frac{P(H_i = 0, T_i = t)}{P(T_i = t)} \\ &= \frac{\pi_0 f_0(t)}{\pi_0 f_0(t) + \pi_1 f_1(t)}, \end{aligned}$$

so  $\frac{\pi_1}{\pi_0} \frac{fdr}{1-fdr}$  is the likelihood ratio  $\frac{f_0}{f_1}$ . By the Neyman-Pearson lemma,  $\Gamma_k$  are a set of uniformly most powerful rejection regions for testing  $f_0$  against  $f_1$ . Assume that for each  $\alpha \in [0, 1]$ , there is a  $k$  such that  $P(T_i \in \Gamma_k | H_i = 0) = \alpha$ ; if not, we can use a randomization argument to get around this. Take any  $\Gamma$ . Pick  $k$  such that  $P(T_i \in \Gamma_k | H_i = 0) = P(T_i \in \Gamma | H_i = 0)$ . Since

$$pFDR(\Gamma) = \frac{\pi_0 P(T_i \in \Gamma | H_i = 0)}{\pi_0 P(T_i \in \Gamma | H_i = 0) + \pi_1 P(T_i \in \Gamma | H_i = 1)}$$

and  $P(T_i \in \Gamma | H_i = 1) \leq P(T_i \in \Gamma_k | H_i = 1)$ , we have  $pFDR(\Gamma) \geq pFDR(\Gamma_k)$ . Similarly  $pMDR(\Gamma) \geq pMDR(\Gamma_k)$ . Hence  $BE(\Gamma) \geq BE(\Gamma_k)$ , so if there is a minimum Bayes error  $\Gamma$ , there is a  $\Gamma_k$  that does at least as well.  $\square$

So which should we use to assess significance, the local  $FDR$  or the q-value? The answer seems to be both, since each has advantages and disadvantages. Storey argues that the q-value better accounts for the multiple hypotheses being tested, and controls the number of false positives while the local  $FDR$  does not [9]. Efron, on the other hand, points out that the local  $FDR$  gives information more suitable for case-by-case interpretation. An observation's q-value can be misleading, since it may have a high local  $FDR$  but a relatively low q-value, since  $FDR$  is an average of  $fdr$  over a rejection region [3].

Finally, Storey notes that the estimation approach allows us to use the bootstrap to get standard error estimates and confidence intervals for our  $FDR$  estimates [8]. We can do this in the usual way by resampling the  $T_i$ . Storey also suggests a method for automatically choosing  $\lambda$  to minimize the mean squared error of  $p\hat{FDR}$  [8]. It is a standard bootstrap procedure - the only twist is that to estimate the MSE, as opposed to the variance, we need to know the true  $pFDR$  or at least an unbiased estimate. Storey notes that the conservative bias means that  $E(p\hat{FDR}_\lambda) \geq pFDR$

for all  $\lambda$ , so we can reduce the bias by using  $\min_{\lambda} p\hat{FDR}_{\lambda}$ . Then, if we use  $B$  bootstrap replications, we can estimate  $M\hat{S}E_{\lambda} = \frac{1}{B} \sum_b (p\hat{FDR}_b^* - \min_{\lambda} p\hat{FDR}_{\lambda})^2$ , and minimize this over  $\lambda$ .

This section has shown that if we know the null distribution of the statistics, or equivalently, if we are given the p-values, there is a simple, and in some sense optimal procedure that lets us control the  $FDR$ . In the next section, we will see what happens when the null distribution is also unknown.

### 3. EMPIRICAL NULL METHODS

In a typical hypothesis testing situation, we know the null distribution of the test statistic. From now on, we will assume the theoretical null is  $N(0, 1)$ , and we observe values  $z_i$ . Sometimes, however, the theoretically predicted null does not agree with the data. For example, consider the following example. The first comes from a SNP analysis and has 575 data points, and the second, from a microarray study of heart patients with 20246. Histograms of the two data sets are shown in Figures 3.1 and 3.2, along with a  $N(0, 1)$  curve. The null distribution of the statistics is  $N(0, 1)$ , and we expect non-null values to be far away from 0. We can see from the histograms, however, that the data does not look  $N(0, 1)$  even near the center, where we expect nearly all the values to come from null cases. Scaling the  $N(0, 1)$  curve to match the height of the histogram for the heart data would require that  $\pi_0$  be extremely low, 0.68; for the SNP data we would need  $\pi_0 = 0.57$ . If we used the theoretical null in standard  $FDR$  methods, the results would be very misleading.

Efron gives a few reasons for why the theoretical null might fail in [5]. First, the theoretical null might rely on distributional assumptions that are not fulfilled. The t-test, for example, assumes normal data. We can get around this by using permutation methods, but permutation nulls are usually quite close to theoretical nulls.

Second, there may be unobserved variables that explain part of the variation in the data. For example, in the heart patient study, the race of the subjects is a possible covariate that might affect both gene expression and heart disease status. These variables are extraneous to the comparison at hand, and we would eliminate their effect if we knew about them and had the data. Since we don't, their influence on the data remains. They tend to widen the distribution of the statistics, making the actual null wider than the theoretical null. This is because the unobserved variable, being unrelated to the comparison, adds an extra layer of noise to the data. Efron gives an example using a t-test to illustrate this in [2]. Unobserved variables are a major problem in observational studies.

Third, there may be correlations in the data that are not accounted for. In data that comes from t-tests, like the heart study, correlations in the observations can reduce the effective number of degrees of freedom of the test; when the data are transformed to z-values (theoretically  $N(0, 1)$ ), they are too spread out. More disturbingly, Efron showed that even if all the observations are individually distributed as  $N(0, 1)$ , correlations among the observations can make the marginal distribution substantially wider or narrower [4].

When the theoretical null is unrealistic, it makes sense to estimate the null distribution from the data, yielding an "empirical null." Efron gives two ways to do this in [5]. Both start by estimating the overall density  $f$  of the data. Efron uses

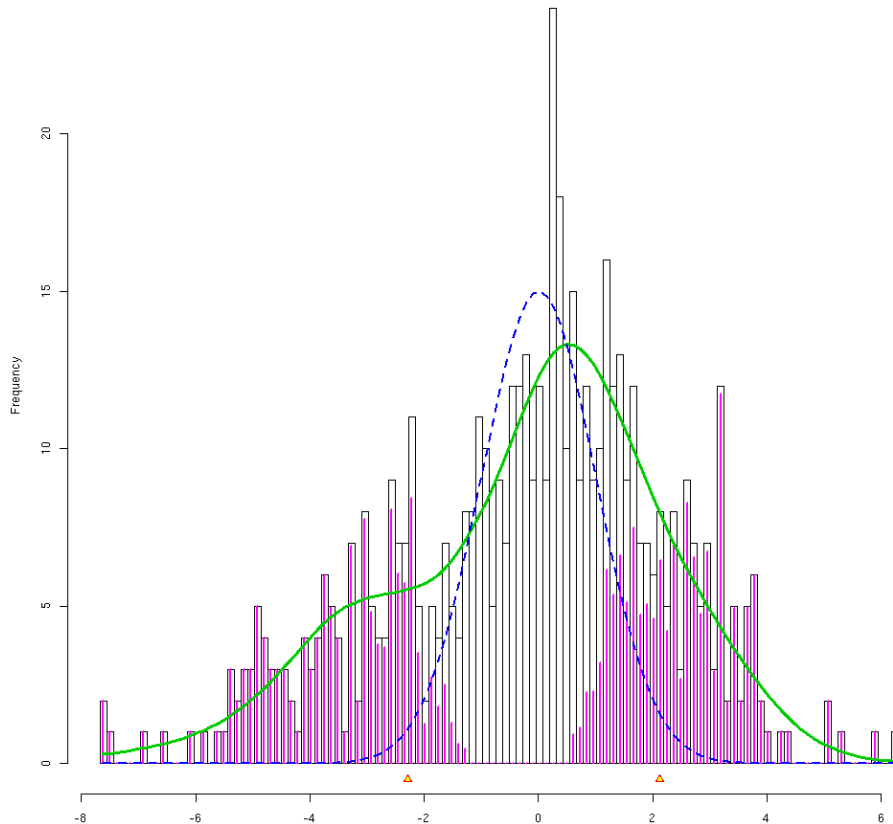


FIGURE 3.1. Histogram of SNP Data. The blue curve is  $N(0, 1)$ . The green curve is the estimated mixture density, the pink bars are estimated non-null counts (using the  $fdr$  at the midpoint) and the arrows show where the estimated  $fdr$  is less than 0.2.

a parametric method that estimates  $f$  by maximum likelihood in an exponential family ( $f = \exp(\sum \beta_j z^j)$ ); a kernel density estimate also seems to work well in practice. He then fits a normal distribution to the data in one of two ways. The first, “geometric,” method fits a quadratic polynomial to  $\log \hat{f}$  near 0. This is because we assume most values near 0 are null, and that the null is close to normal. The second, “analytic,” method assumes that all values in a certain interval, usually about two standard deviations around 0, come from null cases, and fits a normal distribution by maximum likelihood to these values (accounting for the truncation). The interval can be picked to minimize the integrated mean squared error of the estimate among all intervals [10]. Both methods assume that  $\pi_0$ , the proportion of null cases, is unknown, and estimate it. In practice, Efron has found that most of

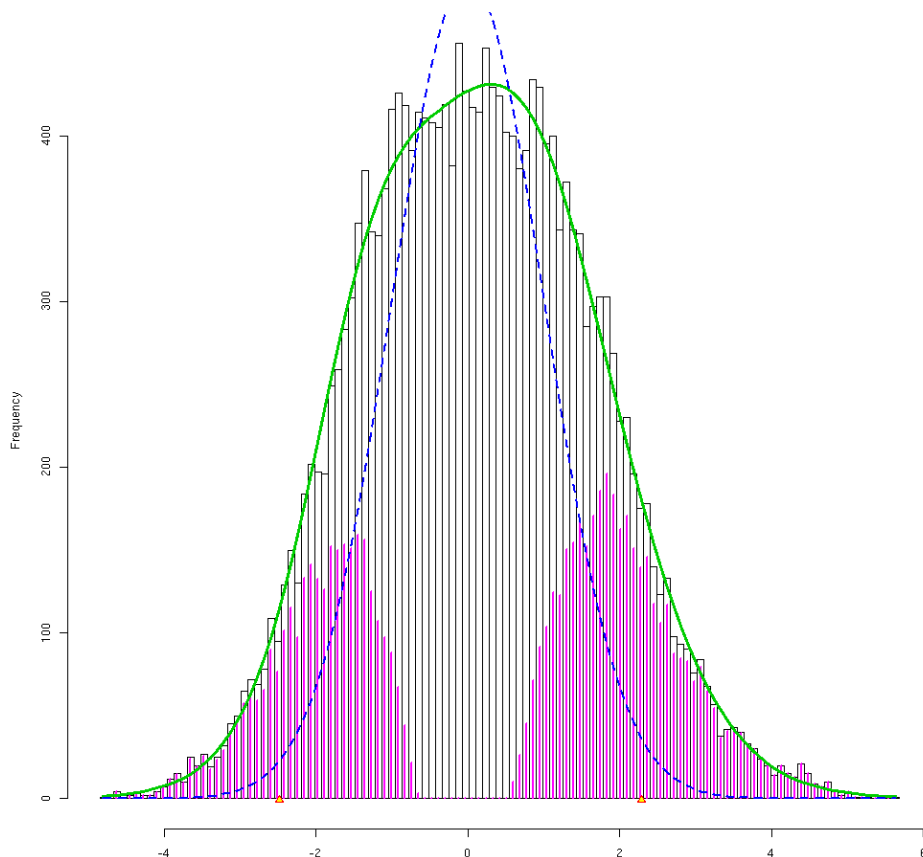


FIGURE 3.2. Histogram of heart patient data. The blue curve is  $N(0, 1)$ . The green curve is the estimated mixture density, the pink bars are estimated non-null counts (using the  $fdr$  at the midpoint) and the arrows show where the estimated  $fdr$  is less than 0.2.

the variability in the  $fdr$  estimates comes from the estimation of the null density, and that the estimation of  $\pi_0$  and  $f$  does not add much noise [3].

The results of these methods are shown in Figures 3.3, 3.4, 3.5 and 3.6. The MLE and geometric fits the data much better than the theoretical  $N(0, 1)$  null. The fits, however, are far from perfect. For the heart data, both the geometric and MLE fits estimate  $\hat{\pi}_0$  to be bigger than 1, which is impossible. The main problem seems to be that the heart data has a wider central peak and thinner tails than we would expect if it were normal - its excess kurtosis is  $-0.254$ . The estimated mixture distribution also suggests that the heart data is distributed asymmetrically.

Skewness is a much bigger problem in the SNP data. The MLE method tries to fit the skewed center of the distribution, and ends up with a very wide null distribution. To compensate for this and match the height of the histogram, it makes  $\hat{\pi}_0$  too large

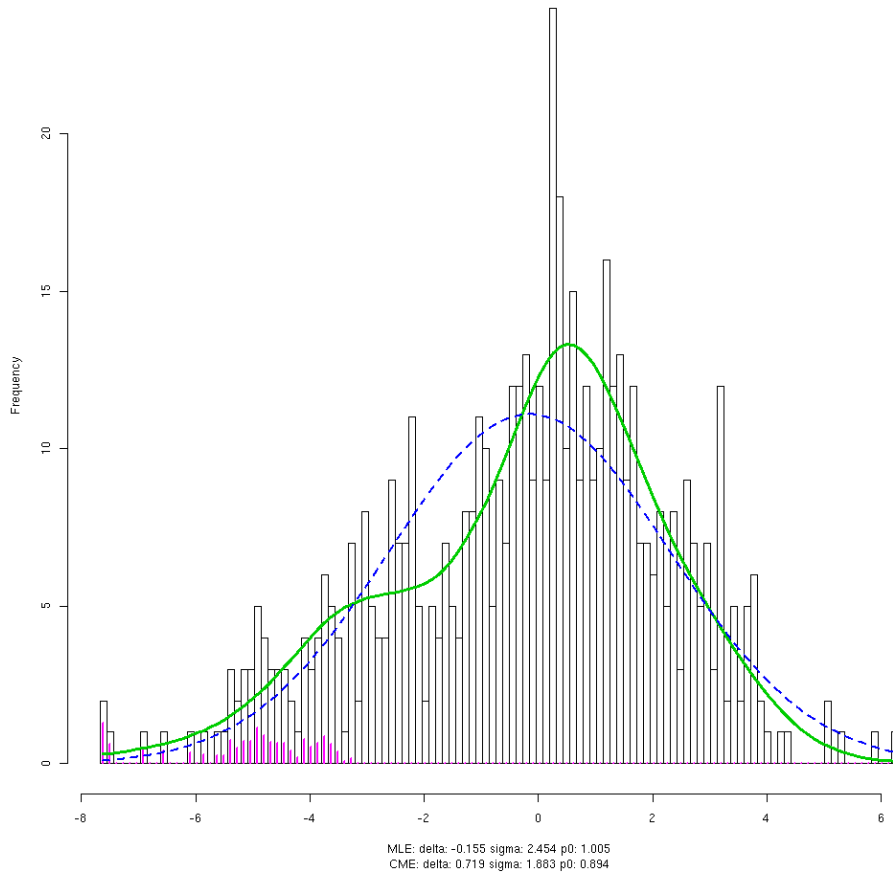


FIGURE 3.3. MLE estimate (blue) of the null for the SNP data. The green curve is the estimated mixture density, the pink bars are estimated non-null counts (using the  $fdr$  at the midpoint) and the arrows show where the estimated  $fdr$  is less than 0.2. , “delta” and “sigma” are the estimated mean and variance of the null, and “p0” is  $\hat{\pi}_0$ . “CME” denotes the estimates for the geometric method.

- bigger than 1. The geometric method gives much more reasonable results in this situation, with  $\hat{\pi}_0 = 0.894$ . We will see later that the behaviors of the MLE and geometric methods here are typical for skewed null distributions.

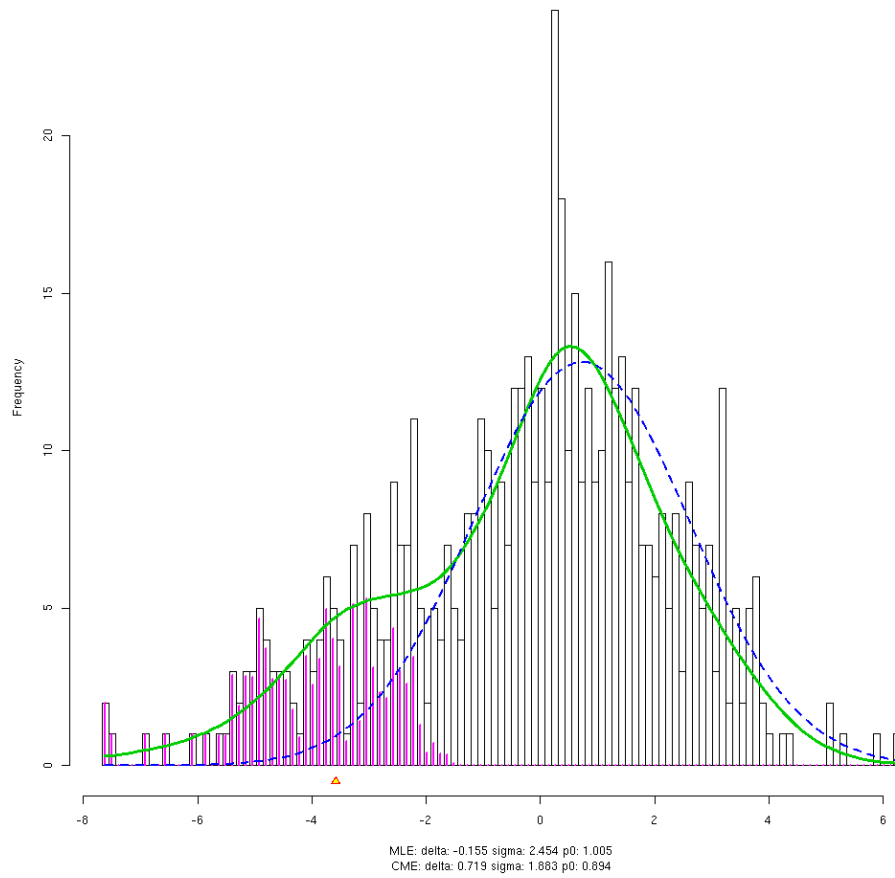


FIGURE 3.4. Geometric estimate (blue) of the null for the SNP data.

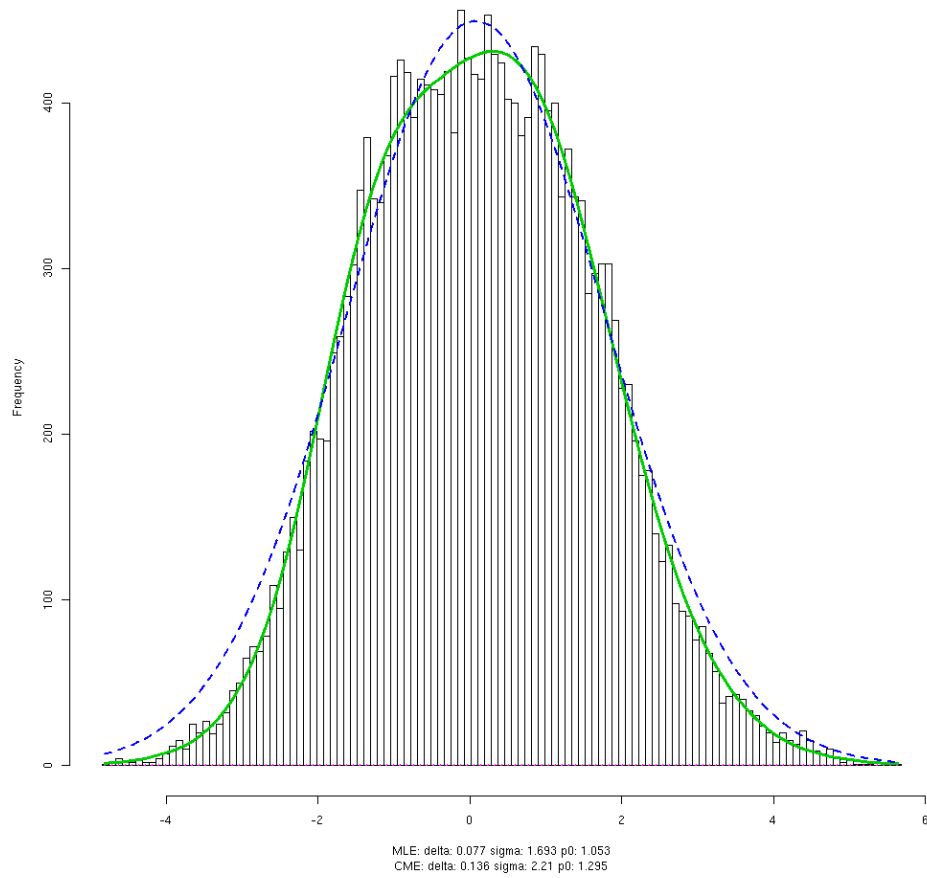


FIGURE 3.5. MLE estimate (blue) of the null for the heart data.

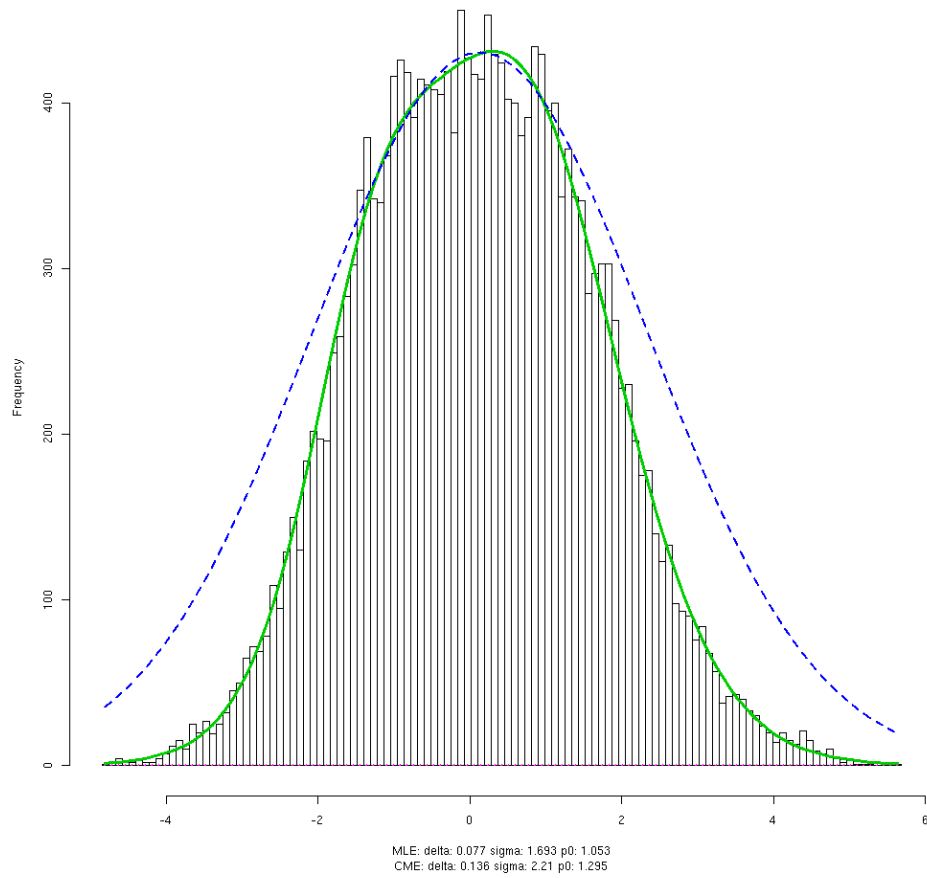


FIGURE 3.6. Geometric estimate (blue) of the null for the heart data.

## 4. ESTIMATING THE NULL USING SYMMETRY

Both of Efron's methods estimate the null by fitting a normal distribution to the center of the data. In some situations, there are other natural estimates of the null density that we can consider. In this section, I will explore a simple method that uses symmetry to estimate the null in certain situations.

Suppose we think that the null distribution is symmetric around 0. In addition, suppose we assume that all the non-null cases will be found to the right. This is a fairly natural situation. For example, if our original data came from a chi-square test of reasonable power, we would expect the non-null cases to be mostly to the right. After we transform the data to be normally distributed (by using  $\Phi^{-1}F_{\chi^2}$ , where  $\Phi$  and  $F_{\chi^2}$  are the normal and chi-square distribution functions), the null distribution will be normal (hence symmetric), and the non-null values will be to the right.

In this situation, the  $z_i$  that are less than 0 all correspond to null cases. This means that the overall density  $f(z)$  agrees with the null density  $f_0(z)$  for  $z < 0$ , and hence our density  $\hat{f}$  is a good estimate of  $f_0$ . By symmetry, however,  $f_0(z) = f_0(-z)$ , so we can just mirror  $\hat{f}$  to get an estimate of  $f_0$ . This gives  $\hat{f}_0(z) = \hat{f}(-|z|)$ . This method can be extended to allow for an arbitrary center of symmetry  $\Delta$  by first estimating  $\Delta$ . Since we expect  $\pi_1 f_1$  to be small compared to  $\pi_0 f_0$  near the center of symmetry, we can estimate  $\Delta$  by a trimmed mean. More simply, if we assume the null density is unimodal, as it typically is, then we could estimate  $\Delta$  by the mode of  $\hat{f}$ . The symmetry method also yields an estimate of  $\pi_0$ . Since we assume the null is symmetric, the total number of null cases is about twice the number less than the center of symmetry, so  $\hat{\pi}_0 = 2 \frac{\#z < \Delta}{\#z}$ .

The symmetry method has a few advantages over fitting a normal to the center. First, it assumes less about the null distribution. In some applications we might expect the null distribution to be symmetric but not normal, perhaps with heavier tails than a normal distribution. In these cases, fitting a normal null will make us underestimate the  $fdr$ , but the symmetry method should still work well.

Second, the symmetry method uses tail information that the central fitting method does not. One drawback of the normal fitting methods is that they rely on the values near the center of the histogram to fit the null distribution. Typically, however, we are most interested in the values the null distribution takes in the tails, since that is where we need  $fdr$  estimates. Since the normal methods only assume that the values near the center are null, they must get information about the tails indirectly, using values at the center and the normality assumption. The symmetry method's assumptions let it use tail data to estimate the tail of the null distribution.

Of course, these advantages rest on the assumptions the symmetry method makes. We will see later that the symmetry method fails badly for asymmetric null distributions. It will also overestimate the  $fdr$  when there are non-null values in the left tail as well as the right. Before that, we can look at the method's performance when its assumptions are fulfilled, and in real data sets where they may or not be.

I first tested the symmetry method using a simple simulation. I used a  $N(0, 1)$  null distribution, a  $N(3, 2)$  non-null, set  $\pi_0 = 0.9$ , used 1000 data points. I assumed the center of symmetry was known. The results are summarized in Table 1. All the methods have positive bias for the most important thresholds 2.5, 3.0, and 3.5,

$z$	$fdr$	Bias			$sd(\log fdr)$			MSE		
		Sym	MLE	Geo	$Sym$	$MLE$	$Geo$	Sym	MLE	Geo
<b>1.5</b>	<b>0.879</b>	0.0265	0.105	-0.0196	<i>0.105</i>	<i>0.0343</i>	<i>0.1349</i>	0.0936	0.1100	0.1118
<b>2.0</b>	<b>0.689</b>	0.0884	0.213	-0.0172	<i>0.193</i>	<i>0.096</i>	<i>0.322</i>	0.164	0.229	0.200
<b>2.5</b>	<b>0.373</b>	0.0942	0.2388	0.0470	<i>0.292</i>	<i>0.234</i>	<i>0.601</i>	0.160	0.273	0.244
<b>3.0</b>	<b>0.124</b>	0.0486	0.1591	0.0621	<i>0.717</i>	<i>0.474</i>	<i>0.922</i>	0.106	0.205	0.195
<b>3.5</b>	<b>0.0288</b>	0.0261	0.0652	0.0346	<i>1.23</i>	<i>0.660</i>	<i>1.26</i>	0.0581	0.0922	0.1142
<b>4.0</b>	<b>0.00545</b>	0.0196	0.0208	0.0166	<i>1.27</i>	<i>0.832</i>	<i>1.63</i>	0.0380	0.0310	0.0783

TABLE 1. Bias, standard deviation of log estimates, and mean squared error for the three empirical null methods.

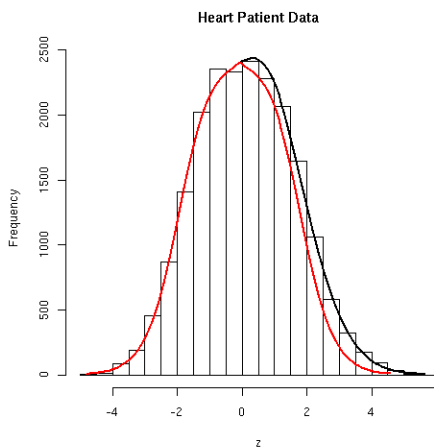


FIGURE 4.1. Symmetry method on heart patient data. The black line is the density estimate, and the red is the symmetry null estimate.

though the geometric method has negative bias for smaller ones. This means that all three methods can be used to control the  $fdr$  in the region where control is most useful. The biases, though, are all a substantial fraction of  $fdr$ , so the methods' power will suffer. The symmetry method is less biased than the MLE method, but is more biased than the geometric. Similarly, its estimates are less variable than the central matching estimates, but are more variable than the MLE estimates. The symmetry method is performing roughly as well as the other two in this case, but is making a different bias-variance tradeoff and uses different assumptions. I then used the symmetry method on the SNP and heart data sets. The fit seems reasonable for each data set, as can be seen in Figures 4.1 and 4.2.

## 5. EVALUATING EMPIRICAL NULL METHODS

Now that we have seen a few different empirical null methods, we might ask how to choose between them, and how to assess their performance more precisely. In this section, I will look at a few ways to do this.

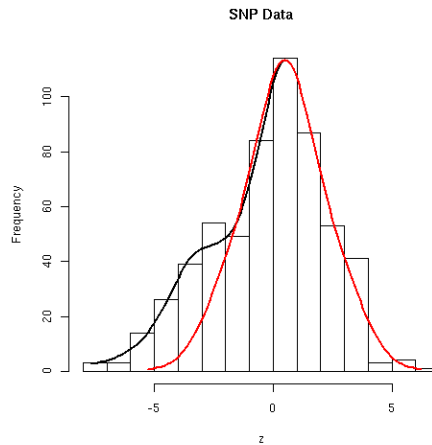


FIGURE 4.2. Symmetry method on SNP data. The black line is the density estimate, and the red is the symmetry null estimate.

The first step is deciding how we will use the method. In this paper, I have looked at empirical null methods in the context of false discovery rates, but they could be useful in other situations. Empirical null methods give us estimates of the null distribution, and knowing the null distribution could have intrinsic value in some settings. A method that produces good overall estimates of the null density may be a poor choice for FDR purposes (for example, because of bad performance in the tails).

Next, we need to consider the type of data. We have already seen an example of this when we compared the symmetry method to normal fitting methods. Choosing a method thus also requires us to specify a range of situations in which we want the method to perform well.

Finally, we need to decide on a measure of the performance of a method. For example, we could look at the mean squared error of  $\hat{fdr}(z)$  for some particular  $z$ , or we could integrate this over the  $z$ 's of interest. These error choices are based on the idea that the empirical null's job is to provide better  $fdr$  estimates. While this is certainly appropriate, it is also useful to think of the whole empirical null and false discovery procedure as a single method to classify cases as null and non-null. We can judge an empirical null method by how it makes the overall method perform.

One way to quantify the performance of the method is by looking at how well it estimates the optimal rejection region. Suppose we have some loss function  $l$  that gives the penalties for misclassification as null or non-null. If we knew the null density and mixture density, we would be able to find the rejection region that minimizes the expected loss  $E_{f_0, f, S}(l) = L(S; f_0, f)$  for a rejection region  $S$ . Call the minimizer  $S^*$ , so  $S^* = \arg \min_S L(S; f_0, f)$ . If we don't know the null density, we can use our empirical null estimate  $\hat{f}_0$  to estimate  $S^*$  by just plugging it into the expected loss function. That is, we estimate  $S^*$  by  $\hat{S}^*$ , where

$$\hat{S}^* = \arg \min_S L(S; \hat{f}_0, f).$$

The distance between  $\hat{S}^*$  and  $S^*$  gives a measure of the performance of our method.

This measure of performance is more natural than the usual criteria of power and  $fdr$  control. Both those quantities are important, but this measure incorporates them by comparing the procedure to the best possible procedure. This performance measure thus accounts for what is achievable in a given setting; power and  $fdr$  control do not. Also, in most cases, we have to make tradeoffs between power and false discoveries. By explicitly bringing these tradeoffs into the loss function, this performance measure shows how well the method performs given the specific tradeoffs we want to make.

We can make this idea more concrete by restricting our attention to situations where we know the form of the family of loss minimizing sets. In particular, if the  $fdr$  is known to decrease in  $z$ , Theorem 2.6 allows us to look only at rejection regions of the form  $[a, \infty)$ . Our loss minimizing procedure reduces to searching for the optimal cutoff value  $c^* = \arg \min_c L(c; f_0, f)$ , where  $L(c; f_0, f) = L([c, \infty); f_0, f)$ . We can summarize the performance of the empirical null method by looking at the squared error  $(\hat{c}^* - c^*)^2$  or any other measure of distance on the real line.

Now that we have a measure of performance, we need to decide on the situations in which we want our method to perform well. One way to generate such a class is to pick a standard situation, and then perturb it in different ways. Here, I will look at skewed versions of the  $N(0, 1)$  and  $N(3, 1)$  example: I will gradually skew the null distribution while keeping the non-null distribution fixed, and see how well the methods adapt to this change.

There are many ways we can skew the normal distribution. One popular choice is to use Edgeworth expansions, but this does not always give a nonnegative density. A more elegant method was given by Azzalini [1]. Let  $\varphi$  and  $\Phi$  be the density and distribution functions of the standard normal, and let  $\alpha \in \mathbb{R}$ . Consider the density

$$\varphi_\alpha(x) = 2\varphi(x)\Phi(\alpha x).$$

As  $\alpha$  varies from  $-\infty$  to  $\infty$ , the density goes from a folded normal on the left, to a left-skewed version of the normal, to a standard normal ( $\alpha = 0$ ), and then correspondingly on the right. A few density plots are shown in Figure 5.1. This family of densities has a nice normality property: if  $X \sim \varphi_\alpha$ , then  $X^2 \sim \chi_1^2$  no matter what  $\alpha$  is. More importantly, this construction gives us an easy way to produce skewed normal distributions that have densities.

To see how the methods performed, I used a skew-normal distribution with location parameter 0 and scale parameter 1 (these are not the mean and variance for the skew normal) for the null, with  $\pi_0 = 0.9$  known. I gave the non-null observations a  $N(3, 1)$  distribution, and considered the mixture density to be known. For my loss function, I used a weighted combination of  $pFDR$  and  $pMDR$ ,  $0.7pFDR + 0.3pMDR$ . I then compared the symmetry method, using the mode to find the center of symmetry, to the MLE and geometric normal fitting methods by seeing how well they estimated the optimal rejection threshold  $c^*$  as  $\alpha$  varied from 0 to 2. The results can be seen in Figures 5.2 and 5.3.

The plots have some expected features. We can see that the symmetry method badly underestimates the best rejection threshold as the null distribution is increasingly skewed to the right. This is what we would have guessed, since as the null is skewed right, the symmetry method sees a smaller and smaller null, since it only looks at the part of the null that is less than the mode of the mixture distribution.

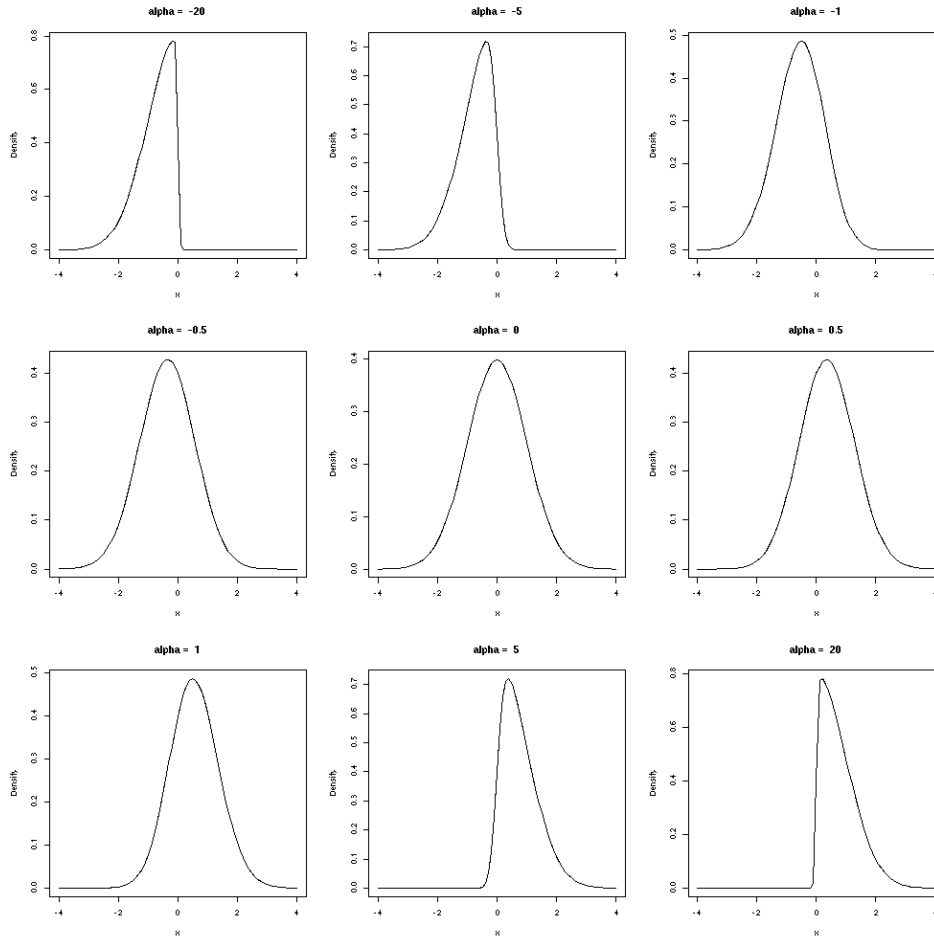


FIGURE 5.1. Plots of the skew normal density for various values of the shape parameter  $\alpha$ .

The poor performance of the MLE is also understandable - the method is maximizing an incorrect likelihood function. Finally, the true best rejection threshold seems to converge to around 4 because as  $\alpha$  gets large, the skew normal converges to a folded normal. The geometric method is clearly the best of the three if we think our null might be skewed; it still gets worse as  $\alpha$  increases, since it becomes narrower to try and match  $\varphi_\alpha$  as it moves to a folded normal. This agrees with what we saw for the SNP data, where the MLE performed poorly but the geometric method gave reasonable results. It is interesting to look at the plots of the achieved losses in Figure 5.3. Although the geometric method estimates the rejection threshold better than the MLE, it has higher loss for most  $\alpha$ . This happens because the geometric method underestimates the best rejection threshold, and this is much more costly than the MLE's overestimation.

It is surprising, though, that the geometric method performs well. We might think that the MLE is the best method to fit a normal, and the geometric method

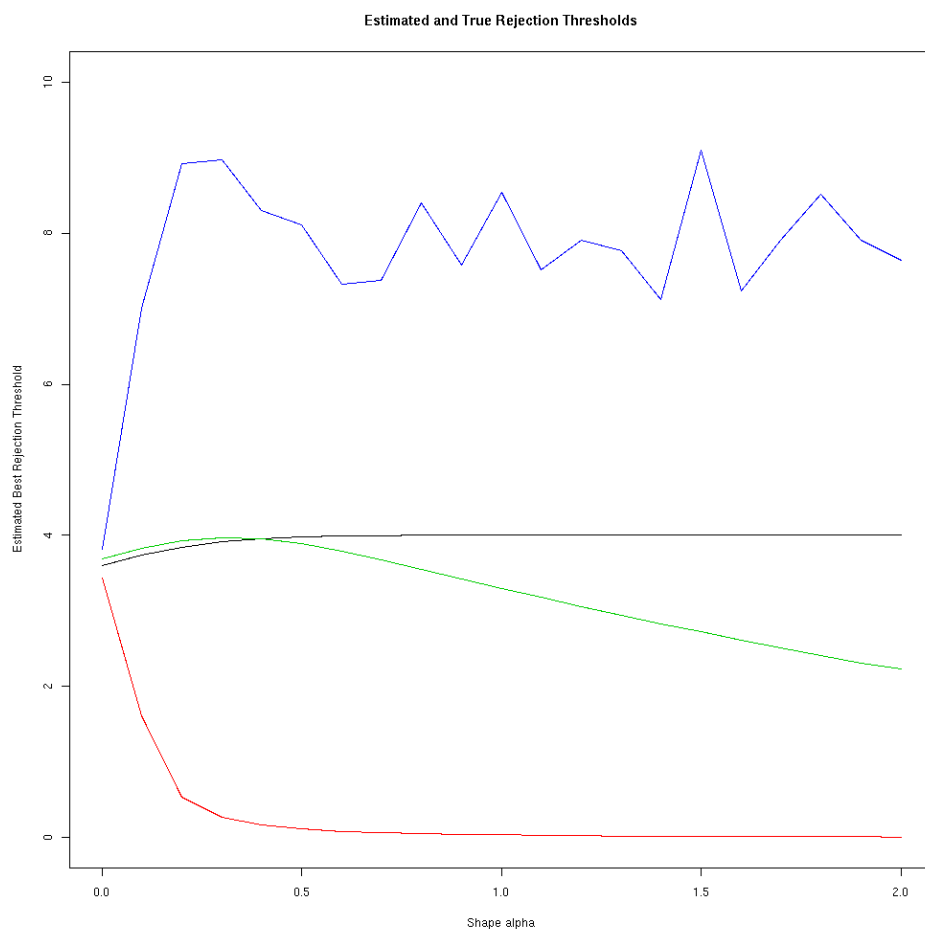


FIGURE 5.2. The true and estimated best rejection thresholds as  $\alpha$  varied from 0 to 2. Black is true, red is symmetry, green is geometric, and blue is MLE.

can only perform worse. It turns out that the MLE fails more badly than we might think here: MLE increases its mean and variance to try and compensate for the skew of the null distribution. The geometric method just fits the center and consequently gives better estimates near  $z = 4$ , where it counts. This is shown in Figure 5.4. Once we see Figure 5.4, however, we might ask why the geometric method outperforms the symmetry method so much, since they seem to give similar density estimates in the critical region near  $z = 4$ . Plotting the estimated cdfs (Figure 5.5) reveals a two things. First, the symmetry method just sees fewer nulls (even though  $\pi_0$  was known, I didn't adjust the symmetry fit to be consistent with this fact since it wasn't clear how that would be done). Second, even though the density estimates look close to the naked eye, this is only because both are so small. We can see this by looking at the shapes of the estimated cdfs - the symmetry cdf levels off much before the geometric cdf. This reinforces the fact that the symmetry cdf is looking

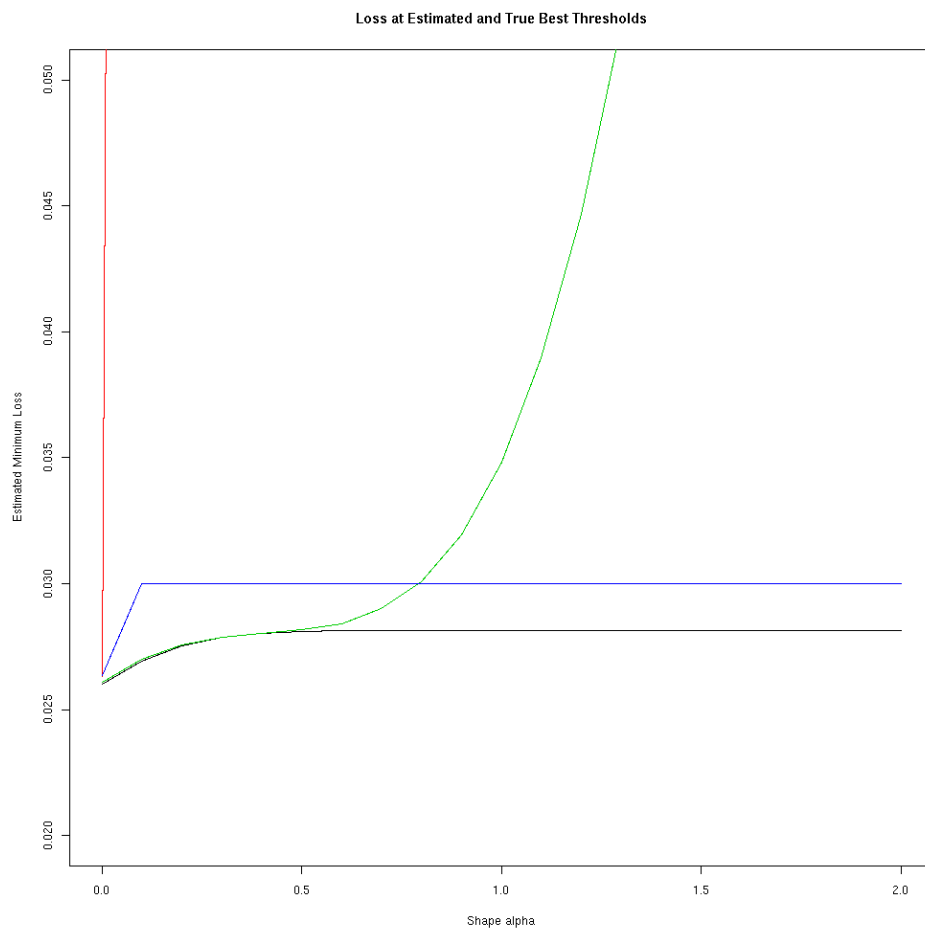


FIGURE 5.3. The true loss at the true and estimated best rejection thresholds as  $\alpha$  varied from 0 to 2. Black is true, red is symmetry, green is geometric, and blue is MLE.

at the left tail of the mixture, instead of the right one, and the ratio of the tails drops to zero (the ratio is  $\frac{\Phi(-\alpha x)}{\Phi(\alpha x)}$ , which converges to 0 as  $x \rightarrow \infty$  for  $\alpha > 0$ ).

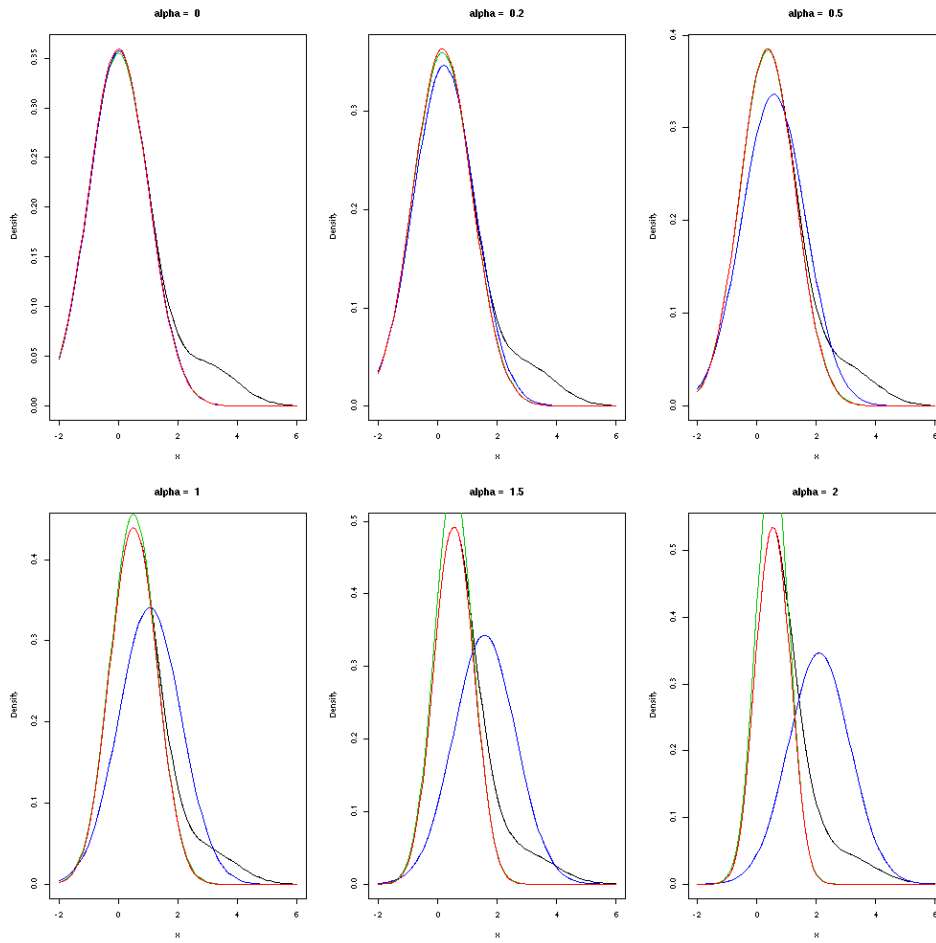


FIGURE 5.4. The mixture density (black) and estimated nulls for various  $\alpha$ . Red is symmetry, green is geometric, and blue is MLE.

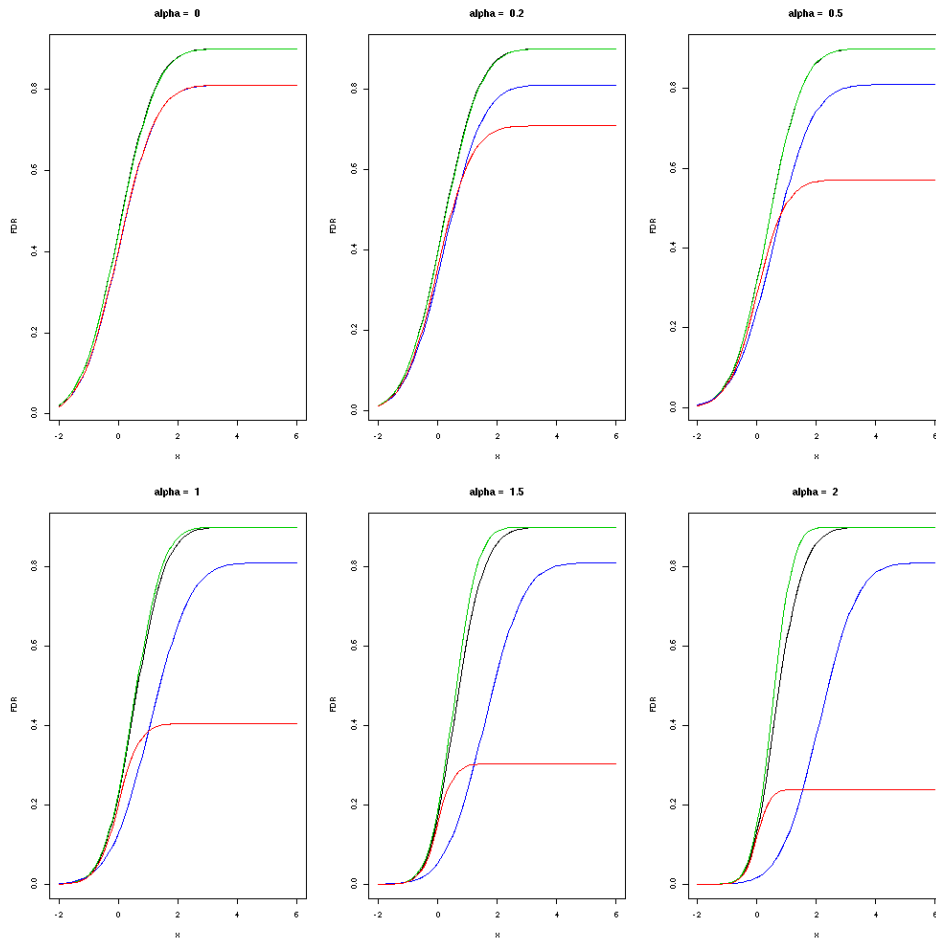


FIGURE 5.5. The null cdf and estimates for various  $\alpha$ . Black is true, red is symmetry, green is geometric, and blue is MLE.

Prior Variance	$\mu$	$\sigma$	$\alpha$
0.1	0.007	1.051	0.025
0.2	0.015	1.046	0.008
0.3	0.010	1.048	0.013
0.4	0.028	1.039	0.004
0.5	0.034	1.049	0.006
0.6	0.028	1.044	0.003
0.7	0.006	1.047	0.002
0.8	0.013	1.041	-0.006
0.9	0.051	1.041	0.001
1.0	0.023	1.042	0.003

TABLE 2. Mean estimates of  $\mu$ ,  $\sigma$  and  $\alpha$  as the variance of the prior  $N(0, \tau)$  on  $\mu$  varies. The priors on  $\sigma$  and  $\alpha$  were  $N(1, 0.2)$  and  $N(0, 0.2)$  respectively.

## 6. BAYESIAN EMPIRICAL NULL METHODS

We might try another approach in our search for optimal empirical null methods. Since Bayesian methods are usually admissible, we can try a more Bayesian approach and see if it reveals anything important about the problem.

One way to do this is to put a prior on the space of distributions and update it with the observed data. The advantage of this approach is that it quantifies our expectation, based on theory, that the data has null distribution  $N(0, 1)$ . The disadvantage is that putting a prior on the space of distributions is quite involved. Here, I'll consider a simple version of this idea. Assume that the null distribution is  $N(0, 1)$ , the non null is  $N(3, 1)$ , and  $\pi_0 = 0.9$ . Suppose we model the null as skew-normal with location  $\mu$ , scale  $\sigma$ , and shape  $\alpha$ , and assume that the central part of the data came from the null. We can priors on the parameters, and then estimate them by their posterior means. To reflect our theoretical expectations, our priors should in some sense center around  $(\mu, \sigma, \alpha) = (0, 1, 0)$ .

For simplicity, I put normal priors on all the parameters (truncated at 0.01 for the variance, with any mass below placed at 0.01), and fit by sampling from the prior distribution and estimating the posterior means. I used 150 data points, and ran the whole process 100 times to see what happened. I then varied the variance of the normal priors to reflect different degrees of confidence in the theoretical null. The results are in Tables 2, 3 and 4. With these priors, the Bayesian method performs well in this situation. I also tried using a scaled  $\chi^2$  prior for  $\sigma$ , and got similar results (Table 5), though the estimates of  $\sigma$  were a bit larger. This is reassuring, since the whole idea behind using an empirical null is to let a large data set speak for itself. We want any Bayesian method to be relatively insensitive to our specific choice of prior.

Unfortunately, this method is poor for a much simpler reason - it doesn't seem to estimate the null very much at all. The prior centering on  $(0, 1, 0)$  caused the posterior estimates to stay quite close to  $(0, 1, 0)$  even when the null distribution was drastically altered. For example, I changed the null to a skew-normal with parameters  $(1, 1.5, 0.2)$  and used priors  $N(0, 1)$ ,  $\frac{1}{4}\chi_4^2$ , and  $N(0, .3)$  for  $\mu$ ,  $\sigma$  and  $\alpha$ . The mean estimate using this method was still  $(0.166, 1.088, 0.036)$ , quite close to

$\tau$	$\mu$	$\sigma$	$\alpha$
0.1	0.026	1.022	-0.003
0.2	0.040	1.040	0.009
0.3	0.027	1.057	0.012
0.4	0.042	1.065	0.014
0.5	0.045	1.077	0.013

TABLE 3. Mean estimates of  $\mu$ ,  $\sigma$  and  $\alpha$  as the parameter  $\tau$  of the truncated  $N(0, \tau)$  prior on  $\sigma$  varies. The priors on  $\mu$  and  $\alpha$  were  $N(0, 0.5)$  and  $N(0, 0.2)$  respectively.

Prior Variance	$\mu$	$\sigma$	$\alpha$
0.1	0.022	1.035	0.001
0.2	0.008	1.044	0.001
0.3	0.021	1.044	0.017
0.4	0.033	1.051	0.023
0.5	0.027	1.042	0.002
0.6	0.004	1.039	0.013
0.7	0.038	1.049	-0.010
0.8	0.018	1.034	0.020
0.9	0.012	1.033	0.004
1.0	0.028	1.047	-0.014

TABLE 4. Mean estimates of  $\mu$ ,  $\sigma$  and  $\alpha$  as the variance of the prior  $N(0, \tau)$  on  $\alpha$  varies. The priors on  $\sigma$  and  $\mu$  were  $N(1, 0.2)$  and  $N(0, 0.5)$  respectively.

$k$	$\mu$	$\sigma$	$\alpha$
3	0.034	1.085	0.010
4	0.045	1.077	0.005
5	0.036	1.070	0.010
6	0.027	1.076	0.005
7	0.041	1.083	0.007
8	0.033	1.070	0.013
9	0.041	1.063	0.000
10	0.046	1.076	0.012

TABLE 5. Mean estimates of  $\mu$ ,  $\sigma$  and  $\alpha$  as the prior  $\frac{1}{k}\chi_k^2$  on  $\sigma$  varies. This prior has mean 1 and variance  $\frac{2}{k}$ . The priors on  $\mu$  and  $\alpha$  were  $N(0, 0.5)$  and  $N(0, 0.2)$  respectively.

the estimates using a  $(0, 1, 0)$  null. Insensitivity to the true null makes this method useless in this situation. This raises a serious doubt about the usefulness of this Bayesian approach, but perhaps a better method based on this idea will perform well.

## 7. SUMMARY

We have seen that false discovery rates and local false discovery rates are powerful tools in multiple hypothesis testing. When the null distribution is known, there are simple procedures that are asymptotically optimal. Often, however, the data do not follow the theoretical null. In this situation, we have seen that estimating the null gives good results. We looked at a few different ways to do this, including a new method that uses symmetry. All of these performed similarly in the simplest case of normal mixtures, with the choice of method corresponding to a bias-variance tradeoff. Next, we looked at a way to evaluate the performance of empirical null methods and saw how well these methods performed when the null distribution was gradually skewed. The symmetry and MLE methods failed badly, but the geometric method did much better. Finally, we tried a simple Bayesian approach to estimating the null, which worked poorly even in a simple test case.

## REFERENCES

1. A. Azzalini and A. Capitanio, *Statistical applications of the multivariate skew-normal distribution*, Journal of the Royal Statistical Society, Series B (Methodological) **61** (1999), 579–602.
2. Bradley Efron, *Large-scale simultaneous hypothesis testing: the choice of a null hypothesis*, JASA **99** (2004), 96–104.
3. ———, *Local false discovery rates*, "http://www-stat.stanford.edu/brad/papers/False.pdf" (2005).
4. ———, *Correlation and large-scale simultaneous significance testing*, "http://www-stat.stanford.edu/brad/papers/Correlation-2006.pdf" (2006).
5. ———, *Microarrays, empirical bayes, and the two-groups model*, "http://www-stat.stanford.edu/brad/papers/twogroups.pdf" (2006).
6. Yoav Benjamini; Yosef Hochberg, *Controlling the false discovery rate: A practical and powerful approach to multiple testing*, Journal of the Royal Statistical Society, Series B (Methodological) **57** (1995), no. 1, 289–300.
7. John Storey; Johnathan Taylor; David Siegmund, *Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: a unified approach*, Journal of the Royal Statistical Society, Series B (Methodological) **66** (2004), no. 1, 187–205.
8. John Storey, *A direct approach to false discovery rates*, Journal of the Royal Statistical Society, Series B (Methodological) **64** (2002), 479–498.
9. ———, *The positive false discovery rate: A bayesian interpretation and the q-value*, The Annals of Statistics **31** (2003), no. 6, 2013–2035.
10. Brit Turnbull, *Optimal estimation of false discovery rates*, Tech. report, Statistics Department, Stanford University, 2007.