Bayesian and Frequentist Multiple Testing

Christopher Genovese[1] and Larry Wasserman[2]
Carnegie Mellon University

April 12, 2002

[DRAFT]

We introduce a Bayesian approach to multiple testing. The method is an extension of the false discovery rate (FDR) method due to Benjamini and Hochberg (1995). We also examine the empirical Bayes approach to simultaneous inference proposed by Efron, Tibshirani, Storey and Tusher (2001). We show that, in contrast to the single hypothesis case – where Bayes and frequentist tests do not agree even asymptotically – in the multiple testing case we do have asymptotic agreement.

KEYWORDS: Multiple Testing, p-values, False Discovery Rate, Bootstrap

[1]Research supported by NSF Grant SES 9866147.

[2]Research supported by NIH Grant R01-CA54852-07 and NSF Grant DMS-98-03433 and NSF Grant DMS-0104016.

1 Introduction

There has been renewed interest in simultaneous inference due to the abundance of large, complex data sets. As a motivating example, consider genetic microarrays. Suppose there are two groups (treatment and control) and m genes. The data, after non-trivial pre-processing, take the following form:

	Control				Treatment			
gene 1	X_{11}	X_{12}	$\cdots$	X_{1k}	Y_{11}	Y_{12}	$\cdots$	Y_{1k}
gene 2	X_{21}	X_{22}	$\cdots$	X_{2k}	Y_{21}	Y_{22}	$\cdots$	Y_{2k}
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$
gene m	X_{m1}	X_{m2}	$\cdots$	X_{mk}	Y_{m1}	Y_{m2}	$\cdots$	Y_{mk}

Each row is a gene and each column is a microarray. Typically, m is around 5,000, and newer arrays will have as many as 50,000 genes. Each data point represents the expression level of the gene. This is a measure of how active the gene is, i.e. how much protein is being produced by that gene. Let $H_i = 1$ if the treatment changes the distribution of the expression level for gene i and $H_i = 0$ otherwise. We want to test

$$H_i = 0 \text{ versus } H_i = 1 \qquad \text{for } i = 1, \ldots, m.$$

The microarray example is a prototype; the ideas that follow apply more generally. The key feature is that m is large but k is small.

2 Modern Frequentist Multiple Testing

Let $P^m = (P_1, \ldots, P_m)$ where P_i is a p-value for testing hypothesis H_i. Multiple testing methods involve choosing a threshold $T = T(P^m)$ and rejecting all hypotheses whose p-values are less than T. The most common methods are uncorrected testing: reject $H_i = 0$ if $P_i < \alpha$; and family-wise corrected testing such as the Bonferroni method: reject $H_i = 0$ if $P_i < \alpha/m$.

Uncorrected testing does not adequately control false positives. Bonferroni and its relatives control the probability of a single error which is too strict

for large m. Two recent methods that avoid these extremes are the *false discovery rate* (FDR) method and the *empirical Bayes testing* (EBT) method. The FDR method is due to Benjamini and Hochberg (1995) – hereafter referred to as BH – and has since been extended by many others including Benjamini and Yekutieli (2002), Storey (2001) and Genovese and Wasserman (2001a, 2001b). The EBT method is due to Efron, Tibshirani, Storey and Tusher (2001). Of course, empirical Bayes is not new but the particular way that Efron et. al. use empirical Bayes for multiple testing has new twists. We would also like to mention that interesting connections between FDR and Bayes are discussed in Storey (2001).

2.1 A Model for Multiple Testing

Let $(P_1, H_1), \ldots, (P_m, H_m)$ be independent pairs such that $P_i \mid H_i = 0 \sim \text{Uniform}(0,1)$ and $P_i \mid H_i = 1 \sim \Xi_i$ for some cdf Ξ_i which is stochastically smaller than the uniform cdf $U(t) = t$. Let $P^m = (P_1, \ldots, P_m)$ and $H^m = (H_1, \ldots, H_m)$. We further assume that $H_i \sim \text{Bernoulli}(a)$ and that the $\Xi_i's$ are random distribution functions drawn from some distribution $\mathcal{L}$. The model may be written as

$$\begin{aligned} H_1, \ldots, H_m &\sim \text{Bernoulli}(a) \\ \Xi_1, \ldots, \Xi_m &\sim \mathcal{L}(d\xi) \\ P_i | H_i = 0, \Xi_i = \xi_i &\sim \text{Uniform}(0,1) \\ P_i | H_i = 1, \Xi_i = \xi_i &\sim \xi_i. \end{aligned}$$

The marginal distribution of P_i is

$$P_1, \ldots, P_m \sim G = (1-a)U + aF \tag{1}$$

where

$$F(t) = \int \xi(t) d\mathcal{L}(\xi). \tag{2}$$

In the nonparametric case we leave $\mathcal{L}$ unspecified and hence F is arbitrary except for the stochastic dominance condition.

The restriction to p-values is not necessary. In general, we let $D_i = (X_{i1}, \ldots, X_{ik}, Y_{i1}, \ldots, Y_{ik})$ denote all the data associated with H_i and we let

V_i be some one-dimensional summary statistic or test statistic derived from D_i. The marginal model for V_i is $R = (1-a)R_0 + aR_1$.

EXAMPLE 2.1 (NORMAL EXAMPLE). *An example that will be useful for illustrative purposes is the following. Let* $V_i \sim N(\theta_i, 1)$ *where* $\theta_i = 0$ *when* $H_i = 0$ *and where* $\theta_i = \theta$ *when* $H_i = 1$. *Here,* θ *is an unknown parameter. Using a common (but unknown) alternative for each hypothesis makes this a toy example.*

2.2 FDR

The Benjamini-Hochberg (BH) procedure rejects all null hypotheses for which $P_i \leq T \equiv P_{(j)}$ where

$$j = \max\left\{0 \leq i \leq m : \ P_{(i)} \leq \alpha \frac{i}{m}\right\}, \tag{3}$$

and $0 \equiv P_{(0)} < P_{(1)} < \cdots < P_{(m)}$ denote the ordered p-values. BH (1995) proved that

$$\mathsf{E}\,(\text{FDR}) \leq (1-a)\alpha \leq \alpha, \tag{4}$$

where FDR is the *realized false discovery rate*, the number of false rejections divided by the number of rejections.[3] The BH result is remarkable: it holds regardless of how many nulls are true and regardless of the distribution of the p-values under the alternatives. In fact, they proved the stronger result

$$\sup_{h^m, \xi^m} \mathrm{E}_{J(h^m, \xi^m)}(\text{FDR}) \leq \alpha$$

where the supremum is over all binary vectors h^m, all vectors of distribution functions $\xi^m = (\xi_1, \ldots, \xi_m)$,

$$J(h^m, \xi^m) = \bigotimes_{i=1}^{m} U^{1-h_i} \xi_i^{h_i}.$$

[3]BH call E(FDR) the false discovery rate. We call FDR the realized false discovery rate and E(FDR) the expected false discovery rate.

There is an alternative way of viewing the Benjamini and Hochberg procedure due to Storey (2002) and Genovese and Wasserman (2001, 2002). Consider rejecting all p-values less than some fixed threshold t. Genovese and Wasserman define the *realized false discovery rate process* by

$$\Gamma(t) = \frac{\sum_i \mathrm{I}(P_i \leq t)(1-H_i)}{\sum_i \mathrm{I}(P_i \leq t) + \prod_i \mathrm{I}(P_i > t)}, \quad 0 \leq t \leq 1 \tag{5}$$

where the second term in the denominator forces $\Gamma(t)$ to be 0 when there are no rejections. Regarded as a function of the threshold t, this is a stochastic process. Typically, $\prod_i \mathrm{I}(P_i > t)$ is exponentially small and hence,

$$\Gamma(t) \approx \frac{\sum_i \mathrm{I}(P_i \leq t)(1-H_i)}{\sum_i \mathrm{I}(P_i \leq t)} \stackrel{d}{=} \frac{\mathrm{Binomial}(m,(1-a)t)}{\mathrm{Binomial}(m,G(t))}.$$

Thus, $\mathrm{E}[\Gamma(t)] = Q(t) + O(m^{-1/2})$ where

$$Q(t) = \frac{(1-a)t}{G(t)}. \tag{6}$$

If we want the expected FDR to be less than α, this suggests choosing a threshold t_* defined by

$$t_* = \max\{t : \ Q(t) \leq \alpha\}. \tag{7}$$

It will then follow that $E(\Gamma(t_*)) = Q(t_*) + O(m^{-1/2}) \leq \alpha + O(m^{-1/2})$. Unfortunately, $t_* \equiv t_*(G,a)$ is a function of G and a which are unknown. An obvious thing to do is to find estimates $\widehat{G}$ and $\widehat{a}$ for G and a and then use the threshold $\widehat{t} = t_*(\widehat{G},\widehat{a})$. To be more explicit,

$$\widehat{t} = \max\{t : \ \widehat{Q}(t) \leq \alpha\}. \tag{8}$$

where

$$\widehat{Q}(t) = \frac{(1-\widehat{a})t}{\widehat{G}(t)}.$$

If we use the conservative estimator $\widehat{a} = 0$, then we get back the Benjamini and Hochberg procedure, as pointed out by Storey (2001). More power can be obtained by taking a less conservative estimator for a such as

$$\widehat{a} = \frac{\widehat{G}(1/2) - \frac{1}{2}}{1 - \frac{1}{2}}$$

which was suggested by Storey (2002). Other estimators of a are considered in Storey (2001) and Genovese and Wasserman (2001b). Since $\widehat{t}$ is obtained from (8) instead of (7), it is not obvious that $E[\Gamma(\widehat{t})] \leq \alpha + O(m^{-1/2})$. After all, $\Gamma(\widehat{t})$ is a random process $\Gamma(\cdot)$ evaluated at a random point $\widehat{t}$ and, moreover, $\Gamma(\cdot)$ and $\widehat{t}$ are correlated. However, Genovese and Wasserman (2001) showed that $E[\Gamma(\widehat{t})] \leq \alpha + O(m^{-1/2})$ does hold assuming reasonable regularity conditions. A close inspection of their proof reveals that this result will continue to hold without the assumption of independence as long as the empirical distribution $\widehat{G}$ is a consistent estimator of G.

The above procedures control the expected FDR. Genovese and Wasserman (2001) introduced *confidence thresholds* which control the realized FDR. Given c and α, a random variable $T = T(P^m)$ is level (c, α) *confidence threshold* if $P(\Gamma(T) < c) \geq 1 - \alpha$. In this paper, we introduce Bayesian procedures that control expected and realized FDR.

On the V scale, assuming that rejections correspond to large values of V, then all the above discussion applies with

$$Q(v) = \frac{(1-a)(1-R_0(v))}{(1-R(v))}. \tag{9}$$

2.3 EBT

The empirical Bayes approach for multiple testing, due to Efron et al (2001), works as follows. First, suppose that a and $f = F'$ are known. Then, Bayes' theorem yields

$$P(H_i = 0|P^m) = P(H_i = 0|P_i) = \frac{1-a}{g(P_i)} \equiv q(P_i) \tag{10}$$

where $g(t) = (1-a) + af(t)$. (The density under the null is $f_0(t) = 1$.) In terms of V, we have $q(V_i) = (1-a)r_0(V_i)/r(V_i)$ where $r_0 = R_0'$ and $r = R'$.

When a and f are not known, we proceed as follows. Let $\widehat{g}$ be an estimate of g. Since $f(t) \geq 0$ for all t, it follows that $g(t) \leq 1 - a$ for all t. Hence, $a \geq 1 - \min g(t)$ which suggests the estimator $\widehat{a} = 1 - \min \widehat{g}(t)$. Finally, define

$$\widehat{q}(P_i) \equiv \frac{1-\widehat{a}}{\widehat{g}(P_i)} = \frac{\min_t \widehat{g}(t)}{\widehat{g}(P_i)}.$$

On the V scale we get $\widehat{a} = 1 - \inf_v \widehat{r}(v)/r_0(v)$ and

$$\widehat{q}(V_i) = \frac{\min_v \frac{\widehat{r}(v)}{r_0(v)}}{\widehat{r}(V_i)}.$$

A similar approach is used in Efron et. al. (2001). Actually, when r_0 and a are unknown, it is no longer the case that H_i is conditionally independent of V_j for $j \neq i$ so one should compute $P(H_i = 0|V^m)$ rather than $P(H_i = 0|V_i)$. However, $P(H_i = 0|V_i)$ can be regarded as an approximation to $P(H_i = 0|V^m)$.

An issue not addressed by Efron et. al. – that we will address in this paper – is the accuracy of $\widehat{q}$. This is important since the cases of interest are when V_i are large and the accuracy of $q(V_i)$ will then be driven by the tails of $\widehat{r}$.

2.4 Thresholds or Posterior Probabilities?

Should we report $P(H_i = 0|V^m)$ or should we choose a threshold T to have a given FDR? Reporting $P(H_i = 0|V^m)$ is informative and intuitive but does not control the number of false positives. On the other hand, choosing an FDR-threshold automatically controls the errors but we lose the quantification of the strength of evidence provided by $P(H_i = 0|V^m)$. Our recommendation is to compute both. This provides the experimenter with the best of both worlds.

The FDR and the posterior probability are related. Consider a fixed threshold $T(P^m) \equiv t$ on the p-value scale. Recall that $E(\Gamma(t)) \approx Q(t) = (1-a)t/G(t)$ and $P(H_i = 0|P^m) = q(t) = (1-a)/g(t)$. Suppose that G is concave, as it is is most problems. Then we have the following FDR-EBT relation:

$$tg(t) \leq \frac{Q(t)}{q(t)} = \frac{tg(t)}{G(t)} \leq 1.$$

Now, consider a Bayesian who chooses to reject whenever $q(t)$ is less than some probability β. This defines a threshold rule $T_B(P^m) = \widehat{q}^{-1}(\beta)$. Under appropriate conditions, $T \xrightarrow{p} q^{-1}(\beta) = t_*$ and thus, asymptotically, $Q(T) \approx Q(t_*) = \beta tg(t)/G(t) \leq \beta$ since $G(t) \geq tg(t)$. From an FDR perspective,

the Bayesian is being conservative. Storey (2001) and Efron et. al. (2001) discuss other relationships between Bayes and FDR. In particular, note that $Q(t) = P(H = 0|P \leq t)$.

EXAMPLE 2.2. *Return to the toy example. Figure 1 shows Q and q and illustrates the conservativeness of the using the posterior probability to define the rejection threshold.*

As we shall see, even if we want to focus on FDR instead of empirical Bayes, we will still need to use the quantity q. Hence, the accuracy of $\widehat{q}$ is of importance from either point of view.

3 Asymptotic behavior of the $\widehat{q}$ process

The Bayesian who wants to report posterior probabilities must use $\widehat{q}$ in place of q since a and g are unknown. Moreover, the confidence threshold methods we describe later also depend on knowing q. The implications of having to estimate q are best understood by examining the asymptotics of $\widehat{q}$ viewed as a stochastic process. For simplicity, first assume that a is known. In what follows we work on the V scale. As a direct consequence of the functional delta method we have the following result.

THEOREM 3.1. *Let $\widehat{r}(v)$ be an estimator of $r(v)$. Suppose that*

$$m^{\alpha}(\widehat{r}(v) - r(v)) \rightsquigarrow W$$

for some $\alpha > 0$, where W is a mean 0 Gaussian process with covariance kernel $\tau(v, w)$. Then

$$m^{\alpha}\ (\widehat{q}(v) - q(v)) \rightsquigarrow Z \tag{11}$$

where

$$Z(v) \stackrel{d}{=} -\frac{(1-a)r_0(v)W(v)}{r^2(v)}.$$

Hence, Z is a Gaussian process with mean 0 and covariance kernel

$$K_q(v, w) = \frac{(1-a)^2\tau(v, w)r_0^2(v)r_0^2(w)}{r(v)^4 r(w)^4}. \tag{12}$$

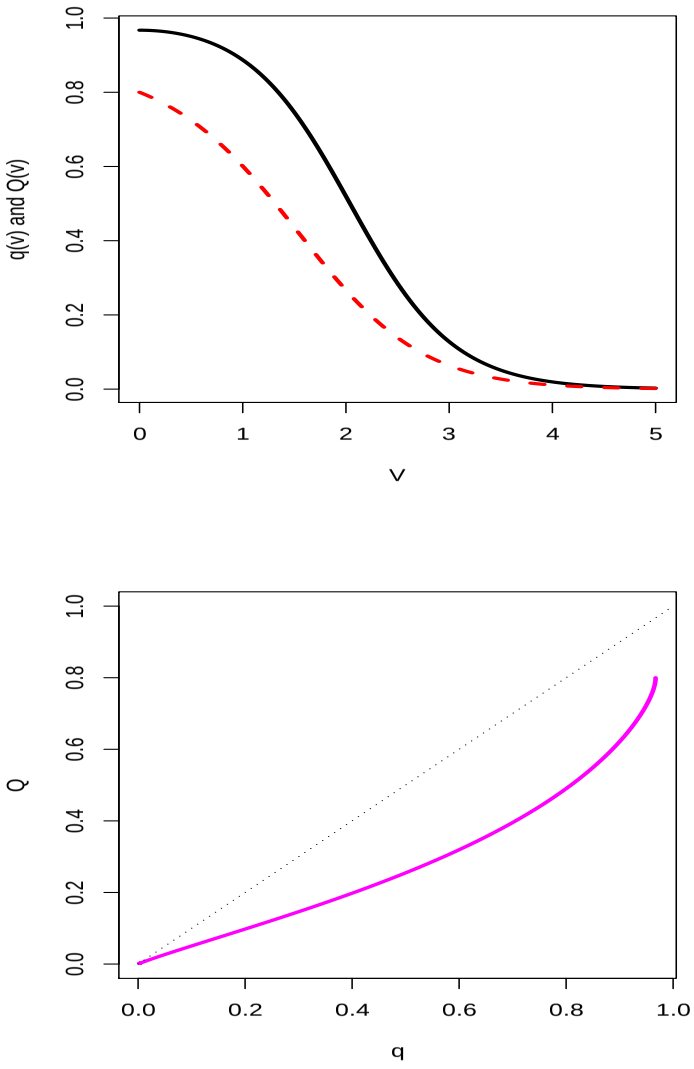

Figure 1: Q and q.

To explore this further, we need to say something specific about $\widehat{r}$. We consider two cases: parametric models and kernel density estimators.

In the parametric case, $q(v) = (1-a)r_0(v)/((1-a)r_0(v)+ar_\theta(v))$ and let us assume that θ is a scalar parameter. Let $\widehat{\theta}$ be a regular, $\sqrt{n}$-consistent estimator of θ and let $\widehat{q}(v) = (1-a)r_0(v)/((1-a)r_0(v)+ar_{\widehat{\theta}}(v))$. The asymptotic standard error of $\widehat{q}(v)$ is

$$\text{se}_v = \text{se}(\widehat{\theta})q(v)(1-q(v))|\dot{\ell}_\theta(v)|$$

where $\text{se}(\widehat{\theta})$ is the standard error of $\widehat{\theta}$ and $\ell_\theta(v) = \log r_\theta(v)$. In the case where $V \sim N(\theta, \sigma^2)$,

$$\text{se}_v = \frac{\sigma}{\sqrt{m}}q(v)(1-q(v))|v-\theta|.$$

Since we are especially interested in cases where $q(v)$ is small, it is more relevant to consider the relative error

$$\text{rel}_v = \frac{\text{se}_v}{q(v)} = \text{se}(\widehat{\theta})(1-q(v))|\dot{\ell}_\theta(v)|.$$

In the Normal case,

$$\text{rel}_v = \frac{\sigma}{\sqrt{m}}(1-q(v))|v-\theta|.$$

In the tails, $1-q(v) \approx 1$ and hence $\text{rel}_v \approx \frac{\sigma}{\sqrt{m}}|v-\theta|$. This suggests that $\widehat{q}(v)$ is reliable in a neighborhood of order $\sqrt{m}$ around θ.

Now consider kernel density estimation:

$$\widehat{r}(v) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{h_m}K\left(\frac{v-V_i}{h_m}\right)$$

where K is a kernel and h_m is the bandwidth. The usual choice of bandwidth $h_m = O(m^{-1/5})$ yields an asymptotically biased estimator and hence (11) fails. Assume, therefore, that we undersmooth the density estimate, for example $h_m = O(m^{-1/4})$. The asymptotic variance of $\widehat{r}(v)$ is $c^2 r(v)/(mh_m)$ where $c^2 = \int K^2(v)dv$. Hence, the standard error $\widehat{q}(v)$ is

$$\text{se}_v = \frac{c\,q(v)}{\sqrt{mh_m r(v)}}$$

with relative error

$$\text{rel}_v = \frac{c}{\sqrt{mh_m r(v)}}.$$

In particular, when $h_m = 1/m^\beta$, the relative error is

$$\text{rel}_v = \frac{c}{m^{(1-\beta)/2}\sqrt{r(v)}}.$$

This will be small when

$$r(v) > \frac{c'}{m^{1-\beta}}$$

for some c', suggesting unreliability in the tails as expected.

EXAMPLE 3.1. *Figure 2 shows the relative error when* $r(v) = .5N(0,1) + .5N(1,1)$. *The solid line is for a kernel density estimator (with* $\beta = 1/4$*) and the dashed line is under the parametric model. The shapes of the curves show the striking difference between the two in the tails.*

When estimation of a is taken into account, things get more complicated. Recall that

$$\widehat{q}(v) = \frac{\min_t \frac{\widehat{r}(t)}{r_0(t)}}{\widehat{r}(v)}.$$

In general, $\widehat{q}$ is not a Hadamard differentiable function of $\widehat{r}$ making it difficult to get a limit law without further assumptions. Results will be reported elsewhere.

4 Bayesian FDR

Let $\Gamma = \{\Gamma(t) : \ 0 \le t \le 1\}$ denote the entire realized FDR process. From the Bayesian point of view, the posterior of Γ is completely determined by the posterior for H^m. Recall that $V_i \sim (1-a)R_0 + aR_1$ where R_1 depends on unknown parameters θ. In the nonparametric case, θ could be infinite dimensional. Let $\psi = (a, \theta)$. Note that the H^m are conditionally independent

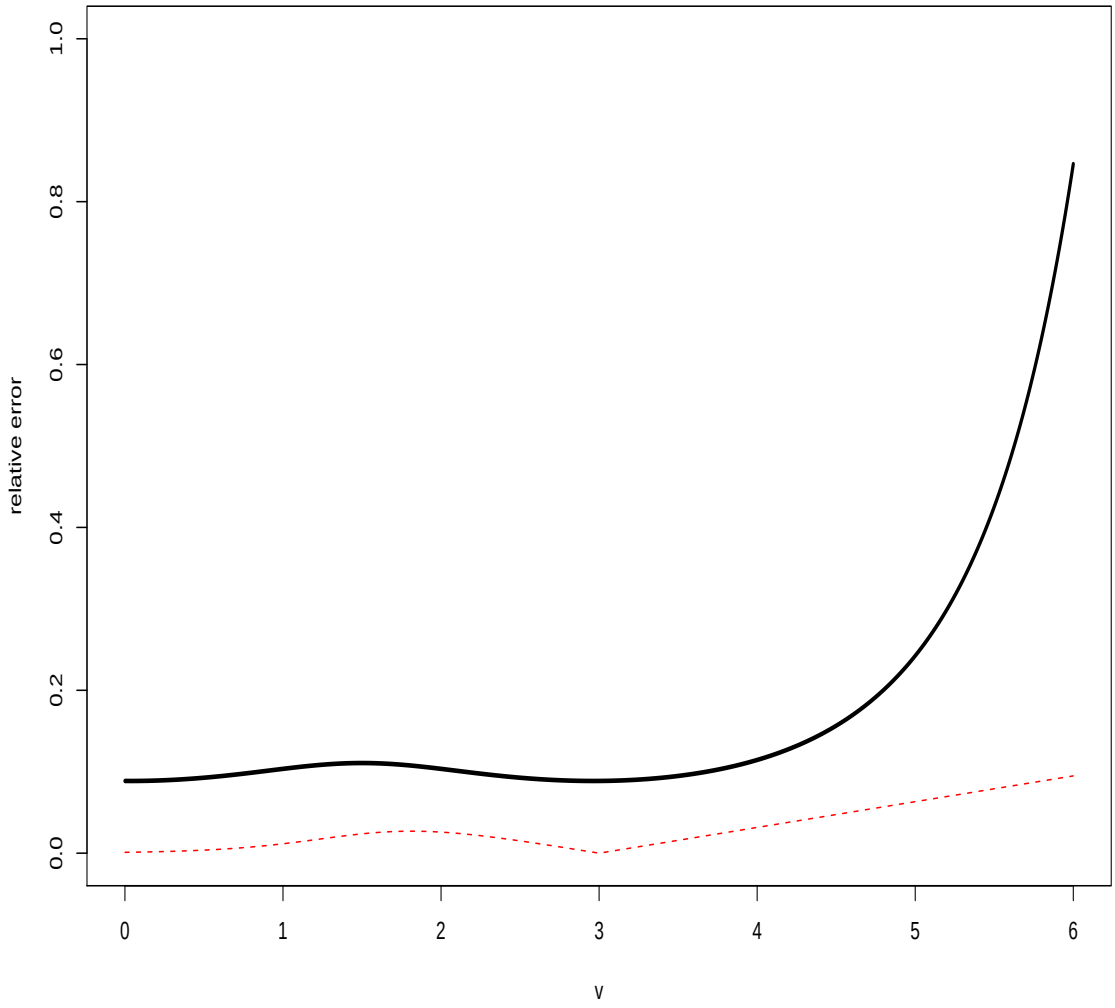

Figure 2: Relative error in estimating $\widehat{q}$. Solid line: nonparametric. Dashed line: parametric.

given V^m and ψ. If V_i is sufficient, H^m are conditionally independent given V^m and ψ. Let $\widehat{\psi}$ be a consistent estimator of ψ. Then,

$$\begin{aligned} P(H^m = h^m|V^m) &= \int P(H^m = h^m|V^m, \psi) f(\psi|V^m) d\psi \\ &\approx P(H^m = h^m|V^m, \widehat{\psi}) \\ &= \prod_i P(H_i = h_i|V_i, \widehat{\psi}) \\ &\approx \prod_i (1 - \widehat{q}(V_i))^{h_i} \widehat{q}(V_i))^{1-h_i}. \end{aligned}$$

It follows that a simple method for generating random draws of Γ from the (approximate) posterior is:

$$\begin{aligned} H_i &\sim \text{Bernoulli}(1 - \widehat{q}(V_i)),\ i = 1, \ldots, m \\ \text{set } \Gamma(v) &= \frac{\sum_i I(V_i > v)(1 - H_i)}{\sum_i I(V_i > v) + \prod_i I(V_i < t)}. \end{aligned}$$

THEOREM 4.1. *Let $\widetilde{\Gamma} \sim \Gamma|V^m$. Suppose that q is known. Under appropriate regularity conditions we have that*

$$\begin{aligned} \widetilde{\Gamma}(v)|V^m &\approx N\left(\frac{\sum_i I(V_i > v) q(V_i)}{\sum_i I(V_i > v)},\ \frac{\sum_i I(V_i > v) q(V_i)(1 - q(V_i))}{(\sum_i I(V_i > v))^2}\right) \\ &\equiv N\left(\mu_q(v), \tau_q^2(v)\right). \end{aligned}$$

When q is unknown, the limiting posterior is a mixture of normals, namely,

$$\widetilde{\Gamma}(v)|V^m \approx \int N\left(\mu_q(v), \tau_q^2(v)\right) d(q|V^m).$$

EXAMPLE 4.1. *Figure 3 is based on 100 observations with $a = .5$ and $\theta = 3$. The first panel shows 1000 draws from the posterior for Γ. The second panel shows $\widehat{q}$ as a function of v. Figure 4 shows the posterior for $\Gamma(2)$ and the Normal approximation for this posterior.*

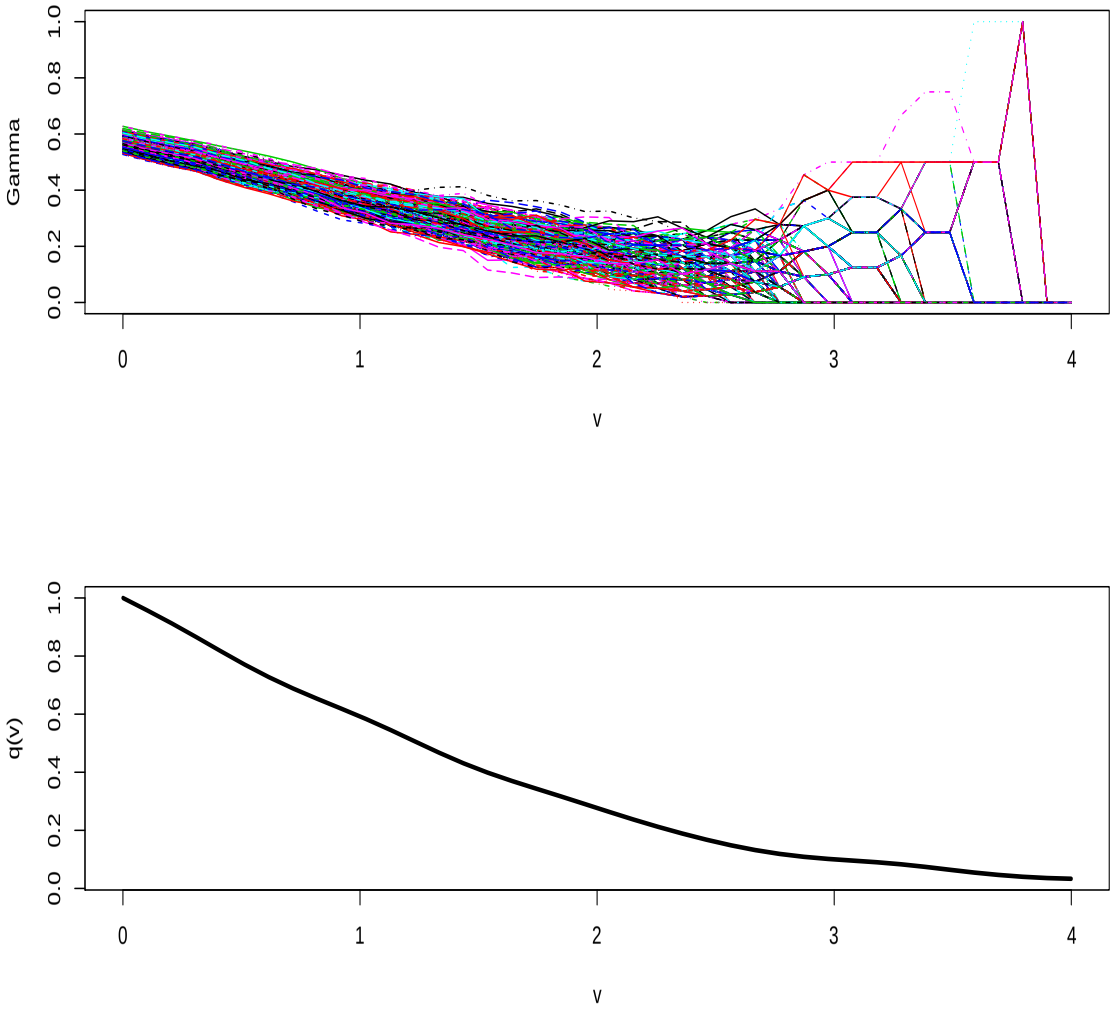

Figure 3: Simulations of Γ from the posterior and $\widehat{q}$.

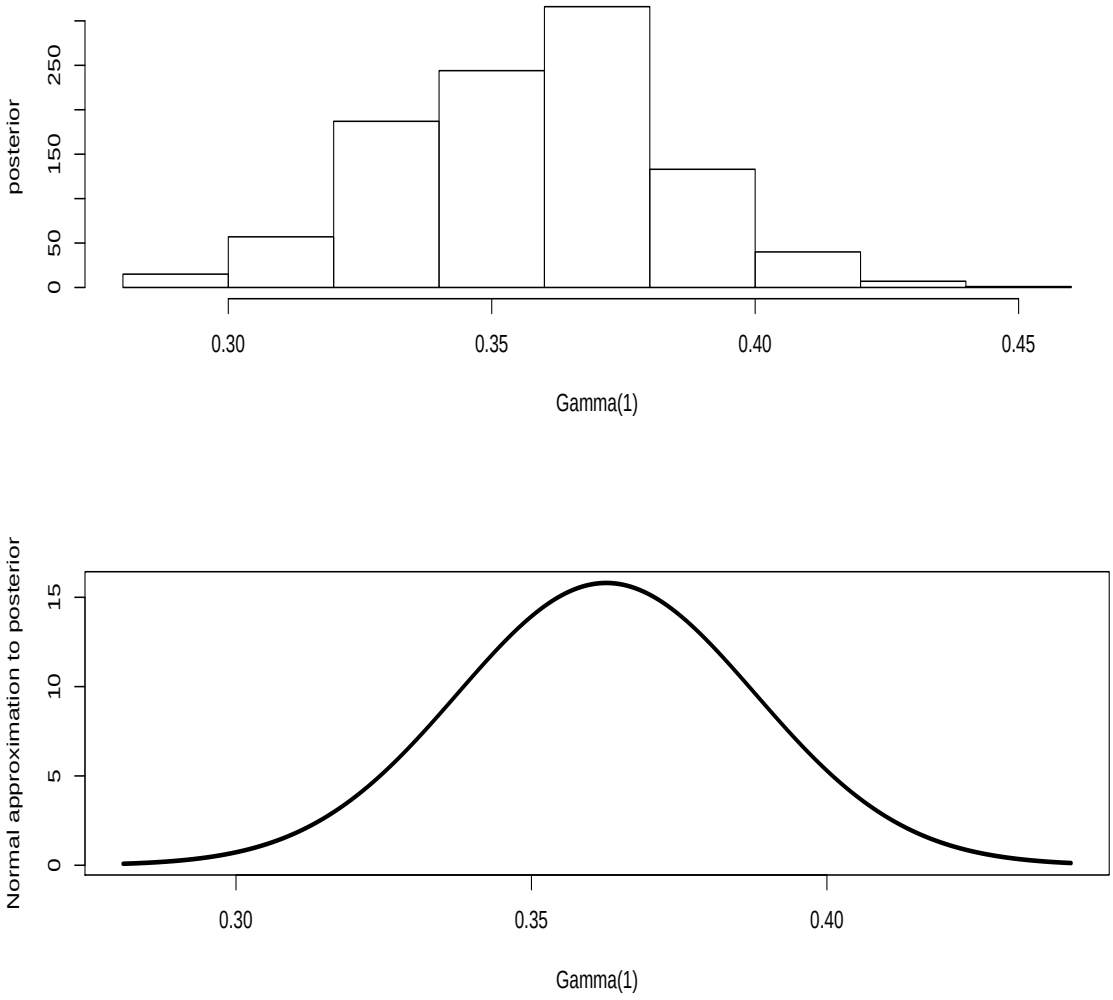

Figure 4: Posterior of $\Gamma(1)$ from simulation and the Normal approximation.

4.1 Controlling Posterior FDR

Rather than controlling $E(FDR)$, a Bayesian would prefer to control E(FDR$|V^m$). Define

$$T_{Bayes} = \sup\{t : \ E(\Gamma(t)|V^m) \leq \alpha\}.$$

By definition, $E(\Gamma(T_{Bayes})|V^m) \leq \alpha\}$ and clearly this rejects as many hypotheses as possible while controlling $E(FDR|V^m)$.

Bayesian confidence thresholds are obtained as follows. Let

$$T = \sup\{t : P(\Gamma(t) > c|V^m) < \alpha\}. \tag{13}$$

It follows that $P(\Gamma(T) < c|V^m) \geq 1 - \alpha$ and hence T is a (c, α) posterior confidence threshold. But is T is a frequentist confidence threshold? If so, this represents an important instance of agreement between Bayes and frequentist in a testing scenario. We explore this in the next section.

Example 4.2. *Let $\alpha = .05$. T_{Bayes} can be obtained directly from the posterior simulation and turns out to be 3.08. On the other hand, with $c = .1$ and $\alpha = .05$, the confidence threshold is 4.0.*

5 Bayes-Frequentist Agreement

For parameter estimation, it is well known that Bayes and frequentist inferences agree asymptotically. For example, consider the Welch-Peers (1995) theorem. If θ is scalar and c_n is chosen to satisfy $P(\theta < c_n|D^n) = 1-\alpha$, then $P_\theta(\theta \in (-\infty, c_n]) = 1 - \alpha + O(n^{-1/2})$. Moreover, if the Jeffreys prior is used then $P_\theta(\theta \in (-\infty, c_n]) = 1-\alpha+O(n^{-1})$. Extensions of this result abound. In this sense, Bayesian and frequentist inference have achieved a certain unification. Testing has resisted such unification. However, in the multiple testing case using FDR we have the following.

Theorem 5.1 (Bayes-Frequentist Agreement). *Fix $t > 0$. Let c_m be such that*

$$P(\Gamma(t) \leq c_m|V^m) = 1 - \alpha.$$

Then, $P(\Gamma(t) \leq c_m) = 1 - \alpha + O(m^{-1/2})$.

A stronger result is that the law of the whole process agrees.

THEOREM 5.2. *Let c be any finite constant. $\mathcal{L}(\Gamma|V^m)$ be the law of $\{\Gamma(v) : v \in [-c, c]\}$ under the posterior. Let $\mathcal{L}_P(\Gamma)$ be the frequentist law of $\{\Gamma(v) : v \in [-c, c]\}$ under P. Under appropriate regularity conditions,*

$$d(\mathcal{L}(\Gamma|V^m), \mathcal{L}_P(\Gamma)) = o_P(1)$$

where d is the Prohorov metric.

The latter result guarantees confidence threshold agreement. That is:

THEOREM 5.3. *Let T be defined as in (13). Then,*

$$P(\Gamma(T) \leq c) \geq 1 - \alpha + o_P(1).$$

REMARK: Despite the agreement, there is an interesting difference between the Bayesian and frequentist approaches. Genovese and Wasserman (2001) shows that there are frequentist confidence thresholds that do not require one to estimate q. On the other hand, it appears that the Bayesian is compelled to estimate q. In the nonparametric case, this might be a serious disadvantage for the Bayesian approach.

REMARK: One could argue that the agreement we have shown is not in the same spirit as the disagreements in testing that have been much discussed. Specifically, we have not focused on measures of evidence in favor of or against a hypothesis. Whether the discussion should be framed this way in multiple testing is an interesting question.

6 False Confidence Rates

In some cases, focusing attention on sharp null hypotheses may be inappropriate. Instead, interval nulls or confidence intervals may be more relevant. In the microarray example, we may be interested in genes whose expression

levels have changed by a factor of 2. If θ_i denotes the difference on a log-scale, this means we are interested in genes for which $|\theta_i| > \log 2$.

More generally, suppose that we call an effect θ_i *interesting* if $|\theta_i| > \delta$ for some fixed $\delta > 0$. Let $C_i = (a_i, b_i)$ be a level $1 - \beta$ confidence interval for θ_i. Let us declare that the i^{th} case is *significant* if $C_i \cap (-t, t) = \emptyset$ where t is some threshold to be chosen. Let R_i be the indicator for the event $\{C_i \cap (-t, t) = \emptyset\}$ and let $\Delta = (-\delta, \delta)$. We define the *false confidence rate* to be

$$\Lambda(t) = \frac{\sum_i R_i \, I(\theta_i \in \Delta)}{\sum_i R_i}.$$

We would like to choose T and β so that $E(\Lambda(T)) \leq \alpha$. It seems natural to choose $\beta = \alpha$ which leaves the problem of choosing T. Assume that $\theta_i \sim F$ for some arbitrary F and that $V_i \approx N(\theta_i, \sigma_n)$. Then,

$$\begin{aligned}
\Lambda(t) &= \frac{\sum_i R_i \, I(\theta_i \in \Delta)}{\sum_i R_i} \\
&\approx \frac{\int_{-\Delta}^{\Delta} P(R_i = 1|\theta) dF(\theta)}{P(R_i = 1)} \\
&\leq \frac{\int_{-\Delta}^{\Delta} dF(\theta) \left[1 - \Phi\left(\frac{t-\Delta}{\sigma_n} - z_{\alpha/2}\right)\right]}{P(R_i = 1)} \\
&\leq \frac{\left[1 - \Phi\left(\frac{t-\Delta}{\sigma_n} - z_{\alpha/2}\right)\right]}{P(R_i = 1)} \\
&\approx \frac{\left[1 - \Phi\left(\frac{t-\Delta}{\sigma_n} - z_{\alpha/2}\right)\right]}{\widehat{P}(R_i = 1)} \qquad (14)
\end{aligned}$$

where

$$\widehat{P}(R_i = 1) = \frac{1}{m} \sum_i R_i.$$

We then select T to be the largest value of t for which the right hand side of (14) is less than α. The procedure can also be based on V_i alone without the use of a confidence intervals. Bayesian and frequentist versions of this procedure will be discussed elsewhere.

7 Microarray Example

Data were obtained on 5355 mouse genes from 3T3L1 (fat) cells over 24 hours after application of Troglitazone which is used to treat diabetes and obesity. The experiment was carried out by Dave Peters and Rob O'Doherty at the University of Pittsburgh. A full analysis of the data will be reported by our group elsewhere. For each gene we have 18 measurements over time. For each gene we computed its median expression level over time and recorded the sign of each measurement as being above or below its median. Let V_i be the longest run of 1's or -1's. In this case V has a discrete distribution on $\{1, \ldots, 18\}$. The null r_0 is know exactly. Figure 5 shows the p-values and 100 draws from the posterior of the Γ process. It is interesting that there is a very steep change in the posterior of Γ between $v = 7$ and $v = 9$. This is much easier to visualize from the posterior draws of Γ than from the p-value plot.

8 Conclusion

As experiments and data sets get increasingly complex, simultaneous inference becomes more important and more common. We have discussed several frequentist and Bayesian methods for dealing with multiple testing problems that arise in these settings. We also briefly discussed interval estimation versions. Extensions are underway to deal with various complications. For example, dependence has been addressed in Benjamini and Yekutieli (2002), Storey and Tibshirani (2002) and Farcome, Genovese and Wasserman (2002). However, dependence can be complex and Bayesian hierarchical models may be useful here.

References

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, 289-300.

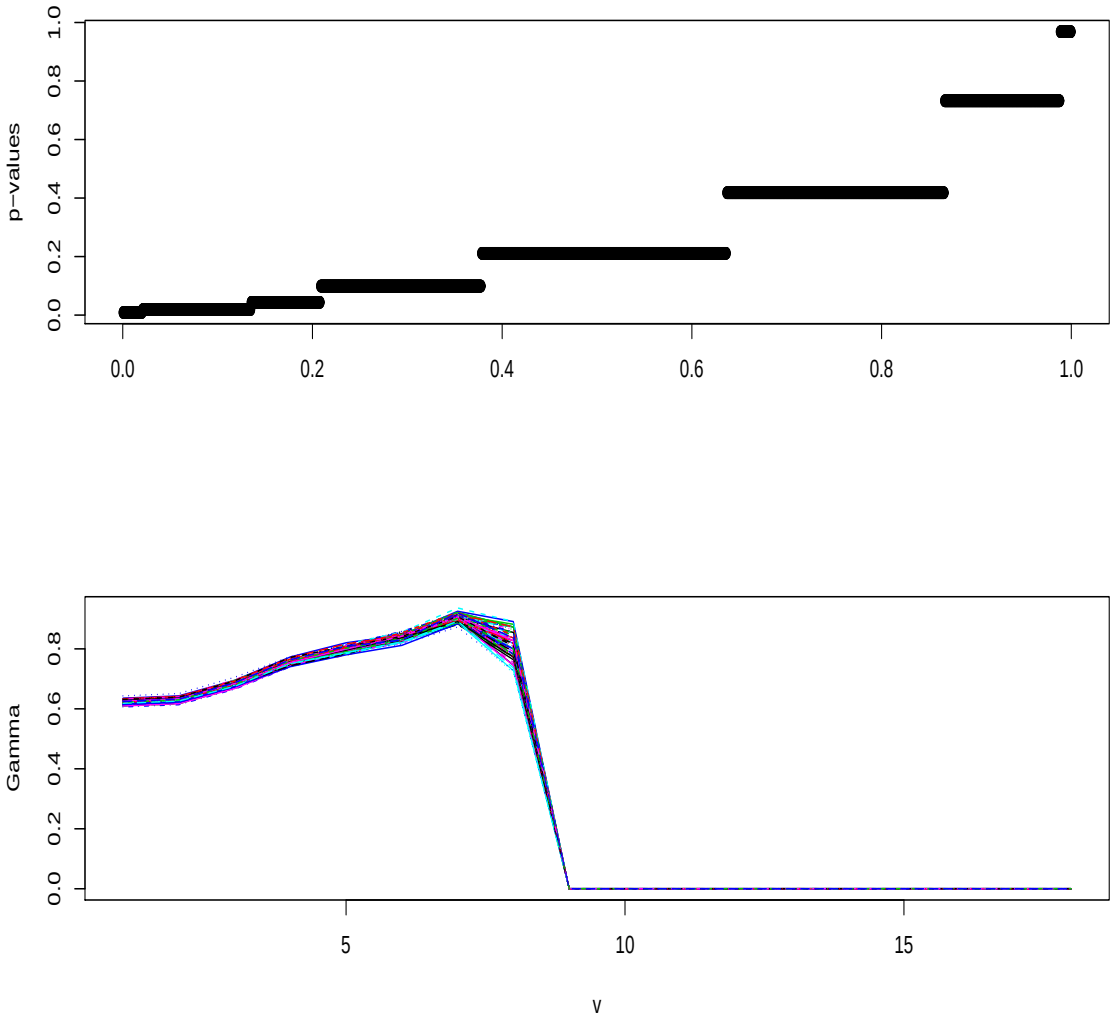

Figure 5: p-value plot and posterior samples of the Γ process.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29, 1165-1188.

Efron, B., Tibshirani R., Storey, J. and Tusher. V. (2001). Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association*, 96, 1151-1160.

Farcome, Genovese and Wasserman (2002). FDR for correlated statistics. In preparation.

Genovese, C. R. and Wasserman, L. (2001). Operating Characteristics and Extensions of the FDR Procedure. *Journal of the Royal Statistical Society B*, to appear.

Genovese, C. R. and Wasserman, L. (2001). False Discovery Rates. Technical report. Department of Statistics. Carnegie Mellon University.

Storey, J. (2001a). A direct approach to false discovery rates. *Journal of the Royal Statistical Society B*, to appear.

Storey, J. (2001b). The positive False Discovery Rate: A Bayesian interpretation and the q-value. Technical Report, Stanford University Dept. of Statistics, `http://www-stat.stanford.edu/ jstorey`.

Storey, J. and Tibshirani, R. (2002). Estimating false discovery rates under dependence with applications to DNA microarrays. Technical report, Stanford.