Tweedie's Formula and Selection Bias

Bradley Efron*
Stanford University

Abstract

We suppose that the statistician observes some large number of estimates z_i, each with its own unobserved expectation parameter μ_i. The largest few of the z_i's are likely to substantially overestimate their corresponding μ_i's, this being an example of *selection bias*, or regression to the mean. Tweedie's formula, first reported by Robbins in 1956, offers a simple empirical Bayes approach for correcting selection bias. This paper investigates its merits and limitations. In addition to the methodology, Tweedie's formula raises more general questions concerning empirical Bayes theory, discussed here as "relevance" and "empirical Bayes information." There is a close connection between applications of the formula and James–Stein estimation.

Keywords: Bayesian relevance, empirical Bayes information, James–Stein, false discovery rates, regret, winner's curse

1 Introduction

Suppose that some large number N of possibly correlated normal variates z_i have been observed, each with its own unobserved mean parameter μ_i,

$$z_i \sim \mathcal{N}(\mu_i, \sigma^2) \qquad \text{for } i = 1, 2, \ldots, N \tag{1.1}$$

and attention focuses on the extremes, say the 100 largest z_i's. *Selection bias*, as discussed here, is the tendency of the corresponding 100 μ_i's to be less extreme, that is to lie closer to the center of the observed z_i distribution, an example of regression to the mean, or "the winner's curse."

Figure 1 shows a simulated data set, called the "exponential example" in what follows for reasons discussed later. Here there are $N = 5000$ independent z_i values, obeying (1.1) with $\sigma^2 = 1$. The $m = 100$ largest z_i's are indicated by dashes. These have large values for two reasons: their corresponding μ_i's are large; they have been "lucky" in the sense that the random errors in (1.1) have pushed them away from zero. (Or else they probably would not be among the 100 largest.) The evanescence of the luck factor is the cause of selection bias.

How can we undo the effects of selection bias and estimate the m corresponding μ_i values? An empirical Bayes approach, which is the subject of this paper, offers a promising solution. Frequentist bias-correction methods have been investigated in the literature, as in Zhong and Prentice (2008, 2010), Sun and Bull (2005), and Zollner and Pritchard (2007). Suggested

*Research supported in part by NIH grant 8R01 EB002784 and by NSF grant DMS 0804324.

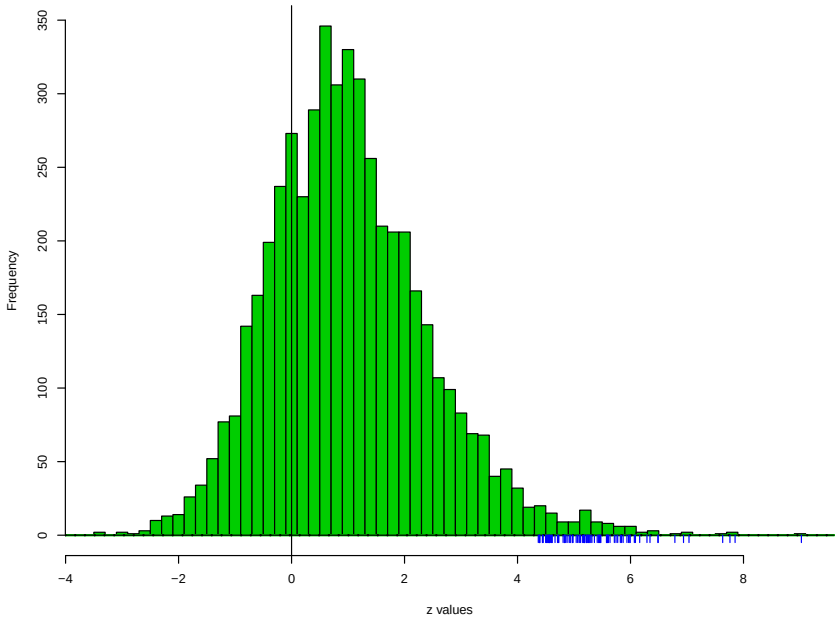

Figure 1: (exponential example) $N = 5000$ z_i values independently sampled according to (1.1), with $\sigma^2 = 1$ and μ_i's as in (3.4); dashes indicate the $m = 100$ largest z_i's. How can we estimate the corresponding 100 μ_i's?

by genome-wide association studies, these are aimed at situations where a small number of interesting effects are hidden in a sea of null cases. They are not appropriate in the more general estimation context of this paper.

Herbert Robbins (1956) credits personal correspondence with Maurice Kenneth Tweedie for an extraordinary Bayesian estimation formula. We suppose that μ has been sampled from a prior "density" $g(\mu)$ (which might include discrete atoms) and then $z \sim \mathcal{N}(\mu, \sigma^2)$ observed, σ^2 known,

$$\mu \sim g(\cdot) \quad \text{and} \quad z|\mu \sim \mathcal{N}(\mu, \sigma^2). \tag{1.2}$$

Let $f(z)$ denote the marginal distribution of z,

$$f(z) = \int_{-\infty}^{\infty} \varphi_\sigma(z-\mu) g(\mu)\, d\mu \qquad \left[\varphi_\sigma(\mu) = (2\pi\sigma^2)^{-1/2} \exp\{-z^2/\sigma^2\}\right]. \tag{1.3}$$

Tweedie's formula calculates the posterior expectation of μ given z as

$$E\{\mu|z\} = z + \sigma^2 l'(z) \qquad \text{where } l'(z) = \frac{d}{dz} \log f(z). \tag{1.4}$$

The formula, as discussed in Section 2, applies more generally — to multivariate exponential families — but we will focus on (1.4).

The crucial advantage of Tweedie's formula is that it works directly with the marginal density $f(z)$, avoiding the difficulties of deconvolution involved in the estimation of $g(\mu)$. This is a great convenience in theoretical work, as seen in Brown (1971) and Stein (1981), and is even more important in empirical Bayes settings. There, all of the observations $z_1, z_2, \ldots, z_N$ can be used to obtain a smooth estimate $\hat{l}(z)$ of $\log f(z)$, yielding

$$\hat{\mu}_i \equiv \hat{E}\{\mu_i|z_i\} = z_i + \sigma^2 \hat{l}'(z_i) \tag{1.5}$$

as an empirical Bayes version of (1.4). A Poisson regression approach for calculating $\hat{l}'(z)$ is described in Section 3.

If the $\hat{\mu}_i$ were genuine Bayes estimates, as opposed to empirical Bayes, our worries would be over: Bayes rule is immune to selection bias, as nicely explained in Senn (2008) and Dawid (1994). The proposal under consideration here is the treatment of estimates (1.5) as being cured of selection bias. Evidence both pro and con, but more pro, is presented in what follows.

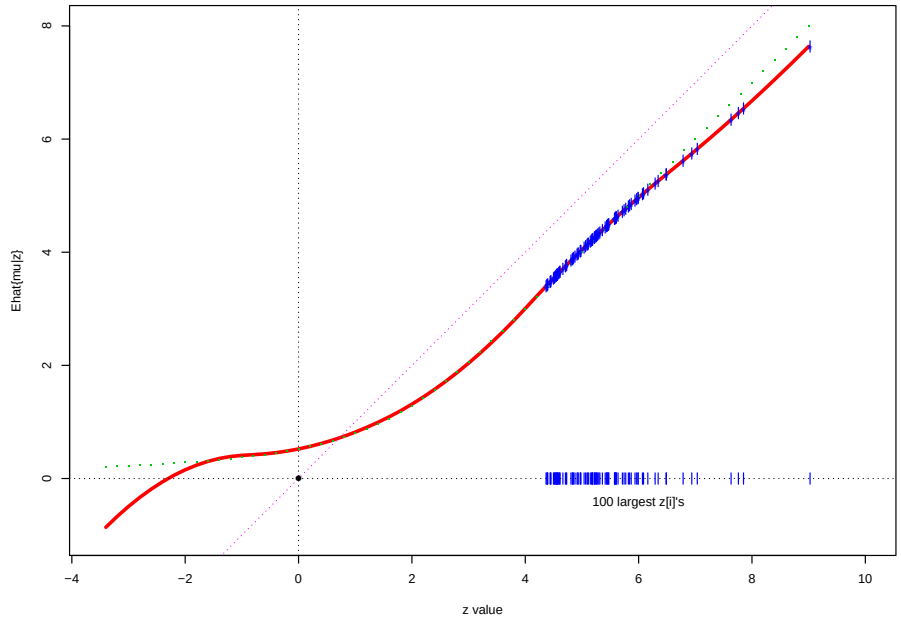

Figure 2: Empirical Bayes estimation curve $\hat{\mu}(z) = z + \hat{l}'(z)$ for the exponential example data of Figure 1, as calculated in Section 3. Dashes indicate the 100 largest z_i's and their corresponding estimates $\hat{\mu}_i$. Small dots show the actual Bayes estimation curve.

Figure 2 graphs $\hat{\mu}(z) = z + \sigma^2\hat{l}'(z)$ ($\sigma^2 = 1$), as a function of z for the exponential example data of Figure 1. For the 100 largest z_i's, the bias corrections $\hat{l}'(z)$ range from -0.97 to -1.40. The $\hat{\mu}_i$'s are quite close to the actual Bayes estimates, and at least for this situation do in fact cure selection bias; see Section 3.

The paper is not entirely methodological. More general questions concerning empirical Bayes theory and applications are discussed as follows: Section 2 and Section 3 concern Tweedie's formula and its empirical Bayes implementation, the latter bringing up a close connection with the James–Stein estimator. Section 4 discusses the accuracy of estimates like that in Figure 2, including a definition of *empirical Bayes information*. A selection bias application to genomics data is presented in Section 5; this illustrates a difficulty with empirical Bayes estimation methods, treated under the name "relevance." An interpretation similar to (1.5) holds for false discovery estimates, Section 6, relating Tweedie's formula to Benjamini and Hochberg's (1995) false discovery rate procedure. The paper concludes in Section 7 with some Remarks, extending the previous results.

2 Tweedie's formula

Robbins (1956) presents Tweedie's formula as an exponential family generalization of (1.2),

$$\eta \sim g(\cdot) \quad \text{and} \quad z|\eta \sim f_\eta(z) = e^{\eta z - \psi(\eta)} f_0(z). \tag{2.1}$$

Here η is the natural or canonical parameter of the family, $\psi(\eta)$ the cumulant generating function or cgf (which makes $f_\eta(z)$ integrate to 1), and $f_0(z)$ the density when $\eta = 0$. The choice $f_0(z) = \varphi_\sigma(z)$ (1.3), i.e., f_0 a $\mathcal{N}(0, \sigma^2)$ density, yields the normal translation family $\mathcal{N}(\mu, \sigma^2)$, with $\eta = \mu/\sigma^2$. In this case $\psi(\eta) = \frac{1}{2}\sigma^2\eta^2$.

Bayes rule provides the posterior density of η given z,

$$g(\eta|z) = f_\eta(z)g(\eta)/f(z) \tag{2.2}$$

where $f(z)$ is the marginal density

$$f(z) = \int_{\mathcal{Z}} f_\eta(z)g(\eta)\, d\eta, \tag{2.3}$$

$\mathcal{Z}$ the sample space of the exponential family. Then (2.1) gives

$$g(\eta|z) = e^{z\eta - \lambda(z)} \left[g(\eta)e^{-\psi(\eta)}\right] \qquad \text{where } \lambda(z) = \log\left(\frac{f(z)}{f_0(z)}\right); \tag{2.4}$$

(2.4) represents an exponential family with canonical parameter z and cgf $\lambda(z)$. Differentiating $\lambda(z)$ yields the posterior cumulants of η given z,

$$E\{\eta|z\} = \lambda'(z), \qquad \text{var}\{\eta|z\} = \lambda''(z), \tag{2.5}$$

and similarly skewness$\{\eta|z\} = \lambda'''(z)/(\lambda''(z))^{3/2}$. The literature has not shown much interest in the higher moments of η given z, but they emerge naturally in our exponential family derivation. Notice that (2.5) implies $E\{\eta|z\}$ is an increasing function of z; see van Houwelingen and Stijnen (1983).

Letting

$$l(z) = \log(f(z)) \quad \text{and} \quad l_0(z) = \log(f_0(z)) \tag{2.6}$$

we can express the posterior mean and variance of $\eta|z$ as

$$\eta|z \sim \left(l'(z) - l_0'(z), l''(z) - l_0''(z)\right). \tag{2.7}$$

In the normal translation family $z \sim \mathcal{N}(\mu, \sigma^2)$ (having $\mu = \sigma^2\eta$), (2.7) becomes

$$\mu|z \sim \left(z + \sigma^2 l'(z), \sigma^2\left(1 + \sigma^2 l''(z)\right)\right). \tag{2.8}$$

We recognize Tweedie's formula (1.4) as the expectation. It is worth noting that if $f(z)$ is *log concave*, that is $l''(z) \le 0$, then var$(\eta|z)$ is less than σ^2; log concavity of $g(\mu)$ in (1.2) would guarantee log concavity of $f(z)$ (Marshall and Olkin, 2007).

The unbiased estimate of μ for $z \sim \mathcal{N}(\mu, \sigma^2)$ is z itself, so we can write (1.4), or (2.8), in a form emphasized in Section 4,

$$E\{\mu|z\} = \text{ unbiased estimate plus Bayes correction.} \tag{2.9}$$

A similar statement holds for $E\{\eta|z\}$ in the general context (2.1) (if $g(\eta)$ is a sufficiently smooth density) since then $-l_0'(z)$ is an unbiased estimate of η (Sharma, 1973).

The $\mathcal{N}(\mu,\sigma^2)$ family has skewness equal zero. We can incorporate skewness into the application of Tweedie's formula by taking $f_0(z)$ to be a standardized gamma variable with shape parameter m,

$$f_0(z) \sim \frac{\text{Gamma}_m - m}{\sqrt{m}} \tag{2.10}$$

(Gamma$_m$ having density $z^{m-1}\exp(-z)/m!$ for $z \geq 0$), in which case $f_0(z)$ has mean 0, variance 1, and

$$\text{skewness } \gamma \equiv 2/\sqrt{m} \tag{2.11}$$

for all members of the exponential family.

The sample space $\mathcal{Z}$ for family (2.1) is $(-\sqrt{m},\infty) = (-2/\gamma,\infty)$. The expectation parameter $\mu = E_\eta\{z\}$ is restricted to this same interval, and is related to η by

$$\mu = \frac{\eta}{1-\frac{\gamma}{2}\eta} \quad \text{and} \quad \eta = \frac{\mu}{1+\frac{\gamma}{2}\mu}. \tag{2.12}$$

Relationships (2.7) take the form

$$\eta|z \sim \left(\frac{z+\gamma/2}{1+\gamma z/2} + l'(z), \frac{1+\gamma^2/4}{(1+\gamma z/2)^2} + l''(z)\right). \tag{2.13}$$

As m goes to infinity, $\gamma \to 0$ and (2.13) approaches (2.8) with $\sigma^2 = 1$, but for finite m (2.13) can be employed to correct (2.8) for skewness. See Remark B, Section 7.

Tweedie's formula can be applied to the Poisson family, $f(z) = \exp(-\mu)\mu^z/z!$ for z a nonnegative integer, where $\eta = \log(\mu)$; (2.7) takes the form

$$\eta|z \sim \big(\text{lgamma}(z+1)' + l'(z),\, \text{lgamma}(z+1)'' + l''(z)\big) \tag{2.14}$$

with lgamma the log of the gamma function. (Even though $\mathcal{Z}$ is discrete, the functions $l(z)$ and $l_0(z)$ involved in (2.7) are defined continuously and differentiably.) To a good approximation, (2.14) can be replaced by

$$\eta|z \sim \left(\log\left(z+\frac{1}{2}\right) + l'(z), \left(z+\frac{1}{2}\right)^{-1} + l''(z)\right). \tag{2.15}$$

Remark A of Section 7 describes the relationship of (2.15) to Robbins' (1956) Poisson prediction formula

$$E\{\mu|z\} = (z+1)f(z+1)/f(z) \tag{2.16}$$

with $f(z)$ the marginal density (2.3).

3 Empirical Bayes estimation

The empirical Bayes formula $\hat{\mu}_i = z_i + \sigma^2\hat{l}'(z_i)$ (1.5) requires a smoothly differentiable estimate of $l(z) = \log f(z)$. This was provided in Figure 2 by means of *Lindsey's method*, a Poisson

regression technique described in Section 3 of Efron (2008a) and Section 5.2 of Efron (2010b). We might assume that $l(z)$ is a Jth degree polynomial, that is,

$$f(z) = \exp\left\{\sum_{j=0}^{J} \beta_j z^j\right\}; \tag{3.1}$$

(3.1) represents a J-parameter exponential family having canonical parameter vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_J)$; β_0 is determined from $\boldsymbol{\beta}$ by the requirement that $f(z)$ integrates to 1 over the family's sample space $\mathcal{Z}$.

Lindsey's method allows the MLE $\hat{\boldsymbol{\beta}}$ to be calculated using familiar generalized linear model (GLM) software. We partition the range of $\mathcal{Z}$ into K bins and compute the counts

$$y_k = \#\{z_i\text{'s in } k\text{th bin}\}, \qquad k = 1, 2, \ldots, K. \tag{3.2}$$

Let x_k be the center point of bin_k, d the common bin width, N the total number of z_i's, and ν_k equal $Nd \cdot f_\beta(x_k)$. The Poisson regression model that takes

$$y_k \overset{\text{ind}}{\sim} \text{Poi}(\nu_k) \qquad k = 1, 2, \ldots, K \tag{3.3}$$

then provides a close approximation to the MLE $\hat{\boldsymbol{\beta}}$, assuming that the N z_i's have been independently sampled from (3.1). Even if independence fails, $\hat{\boldsymbol{\beta}}$ tends to be nearly unbiased for β, though with variability greater than that provided by the usual GLM covariance calculations; see Remark C of Section 7.

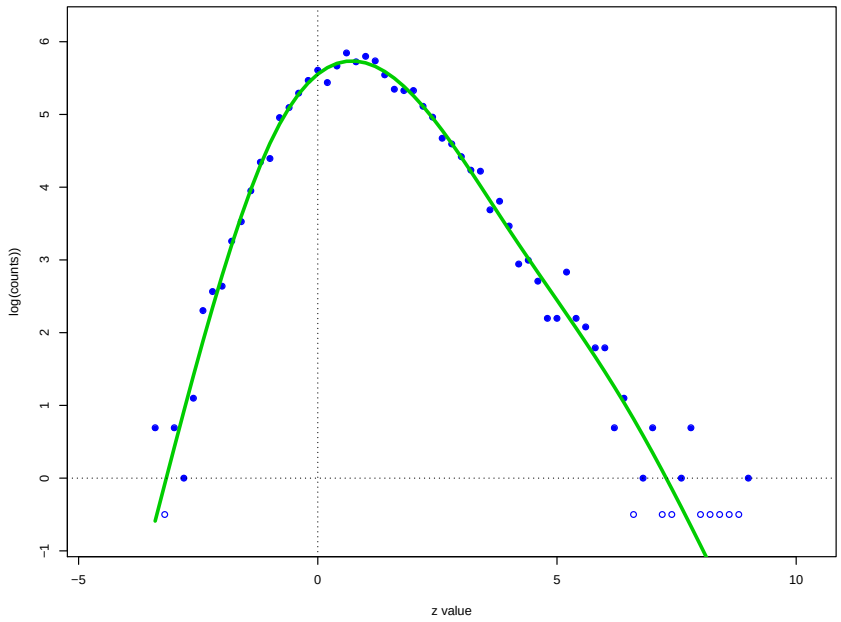

Figure 3: Log bin counts from Figure 1 plotted versus bin centers (open circles are zero counts). Smooth curve is MLE natural spline, degrees of freedom $J = 5$.

There are $K = 63$ bins of width $d = 0.2$ in Figure 1, with centers x_k ranging from -3.4 to 9.0. The bar heights are the counts y_k in (3.2). Figure 3 plots $\log(y_k)$ versus the bin centers x_k.

Lindsey's method has been used to fit a smooth curve $\hat{l}(z)$ to the points, in this case using a natural spline with $J = 5$ degrees of freedom rather than the polynomial form of (3.1), though that made little difference. Its derivative provided the empirical Bayes estimation curve $z+\hat{l}'(z)$ in Figure 2. (Notice that $\hat{l}(z)$ is concave, implying that the estimated posterior variance $1+\hat{l}''(z)$ of $\mu|z$ is less than 1.)

For the "exponential example" of Figure 1, the $N = 5000$ μ_i values in (1.1) comprised 10 repetitions each of

$$\mu_j = -\log\left(\frac{j-0.5}{500}\right), \qquad j = 1, 2, \ldots, 500. \tag{3.4}$$

A histogram of the μ_i's almost perfectly matches an exponential density ($e^{-\mu}$ for $\mu > 0$), hence the name.

Do the empirical Bayes estimates $\hat{\mu}_i = z_i + \sigma^2\hat{l}'(z_i)$ cure selection bias? As a first answer, 100 simulated data sets $\boldsymbol{z}$, each of length $N = 1000$, were generated according to

$$\mu_i \sim e^{-\mu}\ (\mu > 0) \quad \text{and} \quad z_i|\mu_i \sim \mathcal{N}(\mu_i, 1) \qquad \text{for } i = 1, 2, \ldots, 1000. \tag{3.5}$$

For each $\boldsymbol{z}$, the curve $z+\hat{l}'(z)$ was computed as above, using a natural spline model with $J = 5$ degrees of freedom, and then the corrected estimates $\hat{\mu}_i = z_i + \hat{l}'(z_i)$ were calculated for the 20 largest z_i's, and the 20 smallest z_i's. This gave a total of 2000 triples $(\mu_i, z_i, \hat{\mu}_i)$ for the "largest" group, and another 2000 for the "smallest" group.

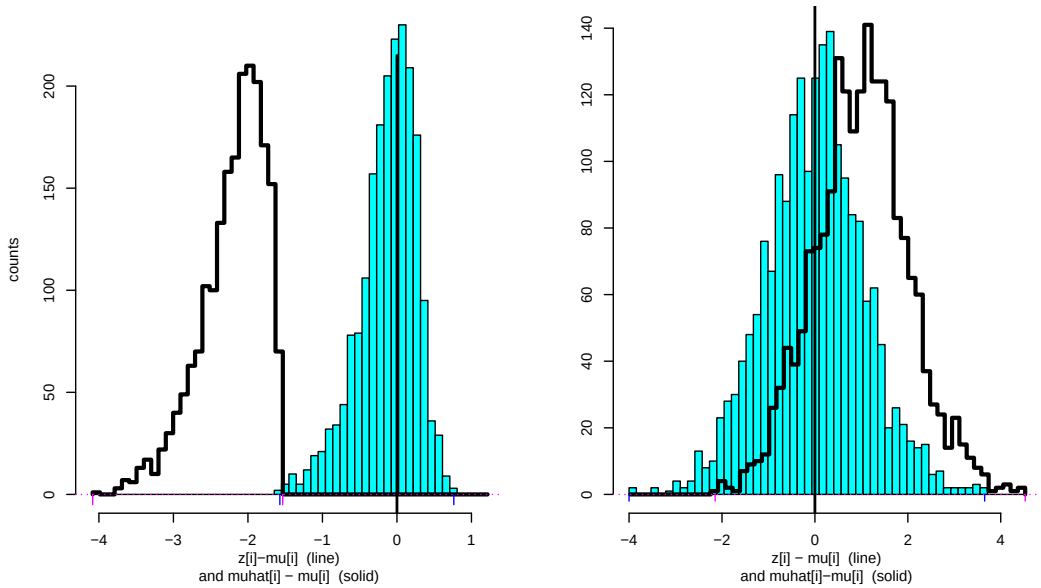

Figure 4: Uncorrected difference $z_i - \mu_i$ (line histograms) compared with empirical Bayes corrected differences $\hat{\mu}_i - \mu_i$ (solid histograms); *left panel* the 20 smallest in each of 100 simulations of (3.5); *right panel* 20 largest, each simulation.

Figure 4 compares the uncorrected and corrected differences

$$d_i = z_i - \mu_i \quad \text{and} \quad \hat{d}_i = \hat{\mu}_i - \mu_i = d_i + \hat{l}'_i \tag{3.6}$$

in the two groups. It shows that the empirical Bayes bias correction $\hat{l}'_i$ was quite effective in both groups, the corrected differences being much more closely centered around zero. Bias

correction usually increases variability but that wasn't the case here, the corrected differences being if anything less variable.

Our empirical Bayes implementation of Tweedie's formula reduces, almost, to the James–Stein estimator when $J = 2$ in (3.1). Suppose the prior density $g(\mu)$ in (1.2) is normal, say

$$\mu \sim \mathcal{N}(0, A) \quad \text{and} \quad z|\mu \sim \mathcal{N}(\mu, 1). \tag{3.7}$$

The marginal distribution of z is then $\mathcal{N}(0, V)$, with $V = A+1$, so $l'(z) = -z/V$ and Tweedie's formula becomes

$$E\{\mu|z\} = (1 - 1/V)z. \tag{3.8}$$

The James–Stein rule substitutes the unbiased estimator $(N-2)/\sum_1^N z_j^2$ for $1/V$ in (3.8), giving

$$\hat{\mu}_i = \left(1 - \frac{N-2}{\sum z_j^2}\right) z_i. \tag{3.9}$$

Aside from using the MLE $N/\sum z_j^2$ for estimating $1/V$, our empirical Bayes recipe provides the same result.

4 Empirical Bayes Information

Tweedie's formula (1.4) describes $E\{\mu_i|z_i\}$ as the sum of the MLE z_i and a Bayes correction $\sigma^2 l'(z_i)$. In our empirical Bayes version (1.5), the Bayes correction is itself estimated from $\boldsymbol{z} = (z_1, z_2, \ldots, z_N)$, the vector of all observations. As N increases, the correction term can be estimated more and more accurately, taking us from the MLE at $N = 1$ to the true Bayes estimate at $N = \infty$. This leads to a definition of *empirical Bayes information*, the amount of information per "other" observation z_j for estimating μ_i.

For a fixed value z_0, let

$$\mu^+(z_0) = z_0 + l'(z_0) \quad \text{and} \quad \hat{\mu}_{\boldsymbol{z}}(z_0) = z_0 + \hat{l}'_{\boldsymbol{z}}(z_0) \tag{4.1}$$

be the Bayes and empirical Bayes estimates of $E\{\mu|z_0\}$, where now we have taken $\sigma^2 = 1$ for convenience, and indicated the dependence of $\hat{l}'(z_0)$ on $\boldsymbol{z}$. Having observed $z = z_0$ from model (1.2), the *conditional regret*, for estimating μ by $\hat{\mu}_{\boldsymbol{z}}(z_0)$ instead of $\mu^+(z_0)$, is

$$\mathrm{Reg}(z_0) = E\left\{\left(\mu - \hat{\mu}_{\boldsymbol{z}}(z_0)\right)^2 - \left(\mu - \mu^+(z_0)\right)^2 \Big| z_0\right\}. \tag{4.2}$$

Here the expectation is over $\boldsymbol{z}$ and $\mu|z_0$, with z_0 fixed. (See Zhang (1997) and Muralidharan (2009) for extensive empirical Bayes regret calculations.)

Define $\delta = \mu - \mu^+(z_0)$, so that

$$\delta|z_0 \sim \left(0, 1 + l''(z_0)\right) \tag{4.3}$$

according to (2.8). Combining (4.1) and (4.2) gives

$$\begin{aligned} \mathrm{Reg}(z_0) &= E\left\{\left(\hat{l}'_{\boldsymbol{z}}(z_0) - l'(z_0)\right)^2 - 2\delta\left(\hat{l}'_{\boldsymbol{z}}(z_0) - l'(z_0)\right) \Big| z_0\right\} \\ &= E\left\{\left(\hat{l}'_{\boldsymbol{z}}(z_0) - l'(z_0)\right)^2 \Big| z_0\right\}, \end{aligned} \tag{4.4}$$

the last step depending on the assumption $E\{\delta|z_0, \boldsymbol{z}\} = 0$, i.e., that observing $\boldsymbol{z}$ does not affect the true Bayes expectation $E\{\delta|z_0\} = 0$. Equation (4.4) says that $\mathrm{Reg}(z_0)$ depends on the squared error of $\hat{l}_{\boldsymbol{z}}(z_0)$ as an estimator of $l'(z_0)$. Starting from models such as (3.1), we have the asymptotic relationship

$$\mathrm{Reg}(z_0) \approx c(z_0)/N \tag{4.5}$$

where is $c(z_0)$ is determined by standard glm calculations; see Remark F of Section 7.

We define the *empirical Bayes information* at z_0 to be

$$\mathcal{I}(z_0) = 1/c(z_0) \tag{4.6}$$

so

$$\mathrm{Reg}(z_0) \approx 1/(N\mathcal{I}(z_0)). \tag{4.7}$$

According to (4.4), $\mathcal{I}(z_0)$ can be interpreted as the amount of information per "other" observation z_j for estimating the Bayes expectation $\mu^+(z_0) = z_0 + l'(z_0)$. In a technical sense it is no different than the usual Fisher information.

For the James–Stein rule (3.8)–(3.9) we have

$$l'(z_0) = -\frac{z_0}{V} \quad \text{and} \quad \hat{l}'_{\boldsymbol{z}}(z_0) = -\frac{N-2}{\sum_1^N z_j^2}\, z_0. \tag{4.8}$$

Since $\sum z_j^2 \sim V\chi^2_N$ is a scaled chi-squared variate with N degrees of freedom, we calculate the mean and variance of $\hat{l}'_{\boldsymbol{z}}(z_0)$ to be

$$\hat{l}'_{\boldsymbol{z}}(z_0) \sim \left(l'(z_0), \frac{2}{N-4}\left(\frac{z_0}{V}\right)^2\right) \tag{4.9}$$

yielding

$$N \cdot \mathrm{Reg}(z_0) = \frac{2N}{N-4}\left(\frac{z_0}{V}\right)^2 \longrightarrow 2\left(\frac{z_0}{V}\right)^2 \equiv c(z_0). \tag{4.10}$$

This gives empirical Bayes information

$$\mathcal{I}(z_0) = \frac{1}{2}\left(\frac{V}{z_0}\right)^2. \tag{4.11}$$

(Morris (1983) presents a hierarchical Bayes analysis of James–Stein estimation accuracy; see Remark I of Section 7.)

Figure 5 graphs $\mathcal{I}(z_0)$ for model (3.5) where, as in Figure 4, $\hat{l}'$ is estimated using a natural spline basis with five degrees of freedom. The heavy curve traces $\mathcal{I}(z_0)$ (4.6) for z_0 between -3 and 8. As might be expected, $\mathcal{I}(z_0)$ is high near the center (though in bumpy fashion, due to the natural spline join points) and low in the tails. At $z_0 = 4$, for instance, $\mathcal{I}(z_0) = 0.092$. The marginal density for model (3.5), $f(z) = \exp(-(z - \frac{1}{2}))\Phi(z-1)$, has mean 1 and variance $V = 2$, so we might compare 0.092 with the James–Stein information (4.11) at $z_0 = 4 - 1$, $\mathcal{I}(z_0) = (V/3)^2/2 = 0.22$. Using fewer degrees of freedom makes the James–Stein estimator a more efficient information gatherer.

Simulations from model (3.5) were carried out to directly estimate the conditional regret (4.4). Table 1 shows the root mean square estimates $\mathrm{Reg}(z_0)^{1/2}$ for sample sizes $N = 125$, 250, 500, 1000, with the last case compared to its theoretical approximation from (4.5). The

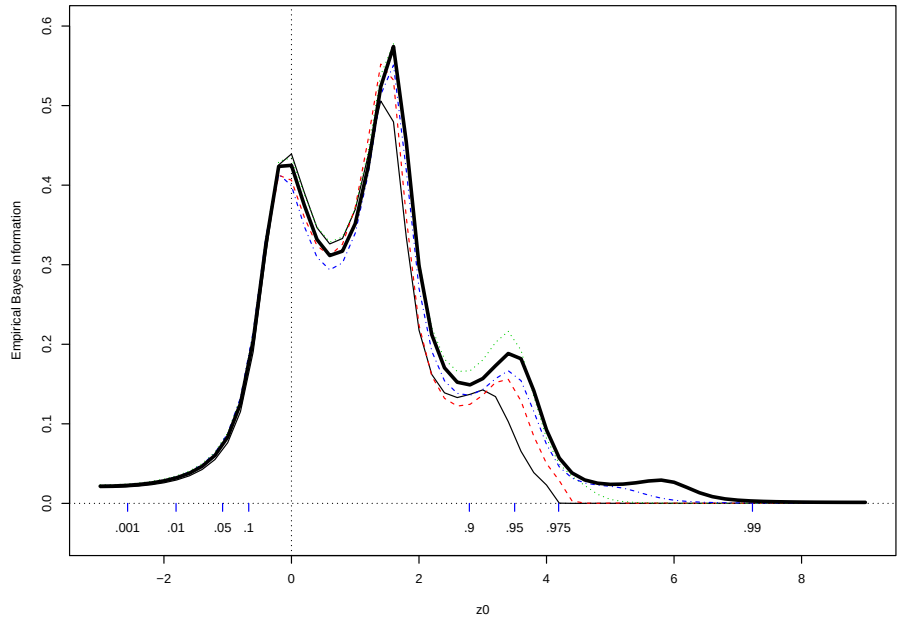

Figure 5: Heavy curve shows empirical Bayes information $\mathcal{I}(z_0)$ (4.6) for model (3.5): $\mu \sim e^{-\mu}$ $(\mu > 0)$ and $z|\mu \sim \mathcal{N}(\mu, 1)$; using a natural spline with 5 degrees of freedom to estimate $\hat{l}'(z_0)$. Light lines are simulation estimates using definition (4.7) and the regret values from Table 1: lowest line for $N = 125$.

theoretical formula is quite accurate except at $z_0 = 7.23$, the 0.999 quantile of z, where there is not enough data to estimate $l'(z_0)$ effectively.

At the 90th percentile point, $z_0 = 2.79$, the empirical Bayes estimates are quite efficient: even for $N = 125$, the posterior rms risk $E\{(\mu - \hat{\mu}(z_0))^2\}^{1/2}$ is only about $(0.92^2 + 0.24^2)^{1/2} = 0.95$, hardly bigger than the true Bayes value 0.92. Things are different at the left end of the scale where the true Bayes values are small, and the cost of empirical Bayes estimation comparatively high.

The light curves in Figure 5 are information estimates from (4.7),

$$\mathcal{I}(z_0) = 1/\left(N \cdot \mathrm{Reg}(z_0)\right) \tag{4.12}$$

based on the simulations for Table 1. The theoretical formula (4.6) is seen to overstate $\mathcal{I}(z_0)$ at the right end of the scale, less so for the larger values of N.

The huge rms regret entries in the upper right corner of Table 1 reflect instabilities in the natural spline estimates of $l'(z)$ at the extreme right, where data is sparse. Some robustification helps; for example, with $N = 1000$ we could better estimate the Bayes correction $l'(z_0)$ for z_0 at the 0.999 quantile by using the value of $\hat{l}'(z)$ at the 0.99 point. See Section 4 of Efron (2009) where this tactic worked well.

5 Relevance

A hidden assumption in the preceding development is that we know which "other" cases z_j are relevant to the empirical Bayes estimation of any particular μ_i. The estimates in Figure 2, for instance, take all 5000 cases of the exponential example to be mutually relevant. Here we will

%ile	0.001	0.01	0.05	0.1	0.9	0.95	0.99	0.999
z_0	-2.57	-1.81	-1.08	-0.67	2.79	3.50	5.09	7.23
$N = 125$	0.62	0.52	0.35	0.22	0.24	0.31	4.60	4.89
250	0.42	0.35	0.23	0.15	0.18	0.17	3.29	6.40
500	0.29	0.24	0.16	0.11	0.11	0.10	0.78	3.53
1000	0.21	0.17	0.11	0.07	0.09	0.08	0.22	1.48
theo1000	**0.21**	**0.18**	**0.12**	**0.08**	**0.08**	**0.07**	**0.20**	**0.58**
$\mathrm{sd}(\mu\|z_0)$	0.24	0.28	0.33	0.37	0.92	0.98	1.00	1.01

Table 1: Root mean square regret estimates $\mathrm{Reg}(z_0)^{1/2}$ (4.4); model (3.5), sample sizes $N = 125$, 250, 500, 1000; also theoretical rms regret $(c(z_0)/N)^{1/2}$ (4.5) for $N = 1000$. Evaluated at the indicated percentile points z_0 of the marginal distribution of z. Bottom row is actual Bayes posterior standard deviation $(1 + l''(z_0))^{1/2}$ of μ given z_0.

discuss a more flexible version of Tweedie's formula that allows the statistician to incorporate notions of relevance.

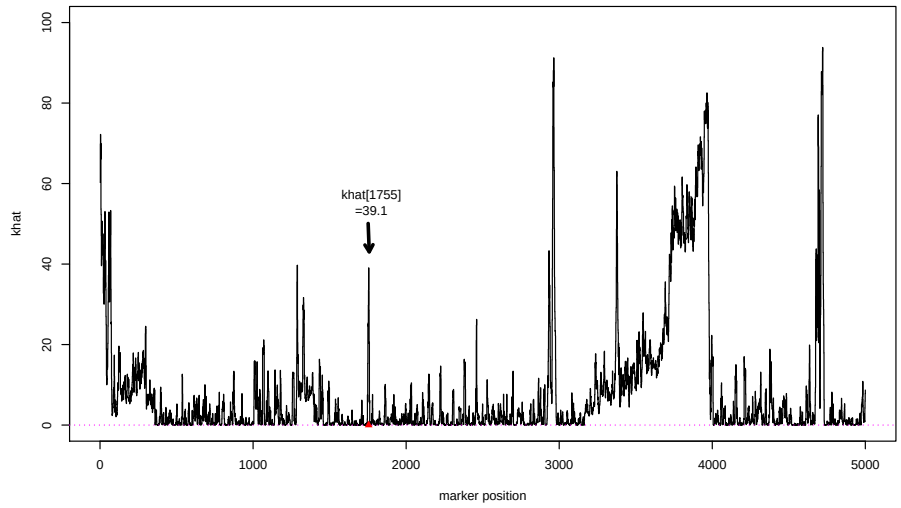

Figure 6: 150 control subjects have been tested for copy number variation at $N = 5000$ marker positions. Estimates $\hat{k}_i$ of the number of cnv subjects are shown for positions $i = 1, 2, \ldots, 5000$. There is a sharp spike at position 1755, with $\hat{k}_i = 39.1$ (Efron and Zhang, 2011).

We begin with a genomics example taken from Efron and Zhang (2011). Figure 6 concerns an analysis of copy number variation (cnv): 150 healthy control subjects have been assessed for cnv (that is, for having fewer or more than the normal two copies of genetic information) at each of $N = 5000$ genomic marker positions. Let k_i be the number of subjects having a cnv at position i; k_i is unobservable, but a roughly normal and unbiased estimate $\hat{k}_i$ is available,

$$\hat{k}_i \doteadot \mathcal{N}(k_i, \sigma^2), \qquad i = 1, 2, \ldots, 5000, \tag{5.1}$$

$\sigma \doteq 6.5$. The $\hat{k}_i$ are not independent, and σ increases slowly with increasing k_i, but we can still

apply Tweedie's formula to assess selection bias. (Remark G of Section 7 analyzes the effect of non-constant σ on Tweedie's formula.)

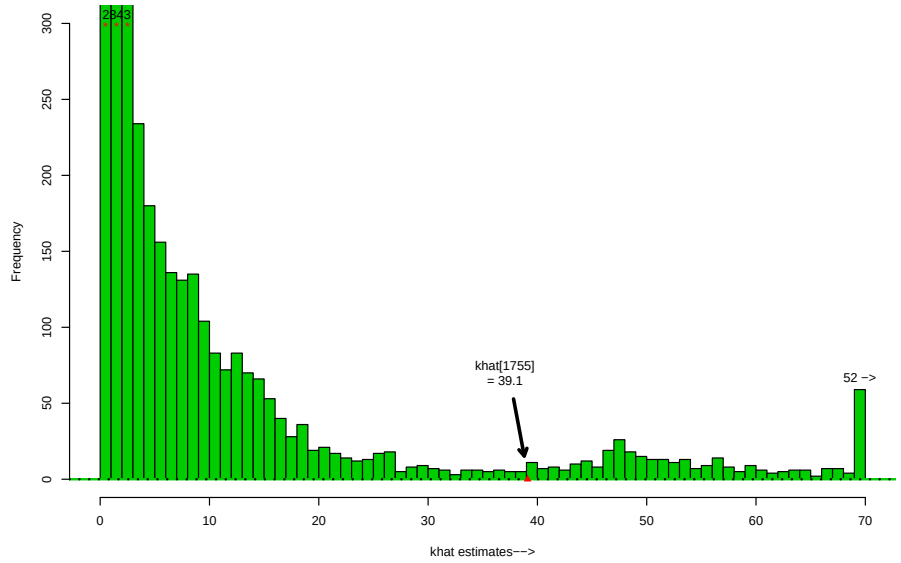

Figure 7: Histogram of estimates $\hat{k}_i,\ i = 1, 2, \ldots, 5000$, for the cnv data; $\hat{k}_{1755} = 39.1$ lies right of the small secondary mode near $\hat{k} = 50$. More than half of the $\hat{k}_i$'s were ≤ 3 (truncated bars at left) while 52 exceeded 70.

The sharp spike at position 1755, having $\hat{k}_{1755} = 39.1$, draws the eye in Figure 6. How concerned should we be about selection bias in estimating k_{1755}? Figure 7 displays the histogram of the 5000 $\hat{k}_i$ values, from which we can carry through the empirical Bayes analysis of Section 3. The histogram is quite different from that of Figure 1, presenting an enormous mode near zero and a small secondary mode around $\hat{k}_i = 50$. However, model (3.1), with $l(z)$ a sixth-degree polynomial, gives a good fit to the log counts, as in Figure 3 but now bimodal.

The estimated posterior mean and variance of k_{1755} is obtained from the empirical Bayes version of (2.8),

$$k_i \big| \hat{k}_i \sim \left(\hat{k}_i + \sigma^2 \hat{l}' \left(\hat{k}_i \right), \sigma^2 \left[1 + \sigma^2 \hat{l}'' \left(\hat{k}_i \right) \right] \right). \tag{5.2}$$

This yields posterior mean and variance $(42.1, 6.8^2)$ for k_{1755}; the upward pull of the second mode in Figure 7 has modestly *increased* the estimate above $\hat{k}_i = 39.1$. (Section 4 of Efron and Zhang (2011) provides a full discussion of the cnv data.)

Relevant cases	1:5000	1:2500	1500:3000	1500:4500
Posterior expectation	**42.1**	**38.6**	**36.6**	**42.8**
Posterior standard deviation	6.8	6.0	6.3	7.0

Table 2: Posterior expectation and standard deviation for k_{1755}, as a function of which other cases are considered relevant.

Looking at Figure 6, one well might wonder if all 5000 $\hat{k}_i$ values are relevant to the estimation

of k_{1755}. Table 2 shows that the posterior expectation is reduced from 42.1 to 38.6 if only cases 1 through 2500 are used for the estimation of $\hat{l}'(k_{1755})$ in (5.2). This further reduces to 36.6 based on cases 1500 to 3000. These are not drastic differences, but other data sets might make relevance considerations more crucial. We next discuss a modification of Tweedie's formula that allows for notions of relevance.

Going back to the general exponential family setup (2.1), suppose now that $x \in \mathcal{X}$ is an observable covariate that affects the prior density $g(\cdot)$, say

$$\eta \sim g_x(\cdot) \quad \text{and} \quad z|\eta \sim f_\eta(z) = e^{\eta z - \psi(\eta)} f_0(z). \tag{5.3}$$

We have in mind a target value x_0 at which we wish to apply Tweedie's formula. In the cnv example, $\mathcal{X} = \{1, 2, \ldots, 5000\}$ is the marker positions, and $x_0 = 1755$ in Table 2. We suppose that $g_x(\cdot)$ equals the target prior $g_{x_0}(\cdot)$ with some probability $\rho(x_0, x)$, but otherwise $g_x(\cdot)$ is much different than $g_{x_0}(\cdot)$,

$$g_x(\cdot) = \begin{cases} g_{x_0}(\cdot) & \text{with probability } \rho(x_0, x) \\ g_{\text{irr},x}(\cdot) & \text{with probability } 1 - \rho(x_0, x), \end{cases} \tag{5.4}$$

"irr" standing for irrelevant. In this sense, $\rho(x_0, x)$ measures the relevance of covariate value x to the target value x_0. We assume $\rho(x_0, x_0) = 1$, that is, that x_0 is completely relevant to itself.

Define R to be the indicator of relevance, i.e., whether or not $g_x(\cdot)$ is the same as $g_{x_0}(\cdot)$,

$$R = \begin{cases} 1 & \text{if } g_x(\cdot) = g_{x_0}(\cdot) \\ 0 & \text{if } g_x(\cdot) \neq g_{x_0}(\cdot) \end{cases} \tag{5.5}$$

and $\mathcal{R}$ the event $\{R = 1\}$. Let $w(x),\ x \in \mathcal{X}$, be the prior density for x, and denote the marginal density of z in (5.3) by $f_x(z)$,

$$f_x(z) = \int f_\eta(z) g_x(\eta)\, d\eta. \tag{5.6}$$

If $f_{x_0}(z)$ were known we could apply Tweedie's formula to it but in the cnv example we have only one observation z_0 for any one x_0, making it impossible to directly estimate $f_{x_0}(\cdot)$. The following lemma describes $f_{x_0}(z)$ in terms of the relevance function $\rho(x_0, x)$.

Lemma 5.1. *The ratio of $f_{x_0}(z)$ to the overall marginal density $f(z)$ is*

$$\frac{f_{x_0}(z)}{f(z)} = \frac{E\{\rho(x_0, x)|z\}}{\Pr\{\mathcal{R}\}}. \tag{5.7}$$

Proof. The conditional density of x given event $\mathcal{R}$ is

$$w(x|\mathcal{R}) = \rho(x_0, x) w(x) / \Pr\{\mathcal{R}\} \tag{5.8}$$

by Bayes theorem and the definition of $\rho(x_0, x) = \Pr\{\mathcal{R}|x\}$; the marginal density of z given $\mathcal{R}$ is then

$$f_{\mathcal{R}}(z) = \int f_x(z) w(x|\mathcal{R})\, dx = \int f_x(z) w(x) \rho(x_0, x)\, dx \Big/ \Pr\{\mathcal{R}\} \tag{5.9}$$

so that

$$\begin{aligned}\frac{f_{\mathcal{R}}(z)}{f(z)} &= \frac{\int f_x(z)w(x)\rho(x_0,x)\,dx}{\Pr\{\mathcal{R}\}f(z)} = \frac{\int \rho(x_0,x)w(x|z)\,dx}{\Pr\{\mathcal{R}\}} \\ &= \frac{E\{\rho(x_0,x)|z\}}{\Pr\{\mathcal{R}\}}.\end{aligned} \tag{5.10}$$

But $f_{\mathcal{R}}(z)$ equals $f_{x_0}(z)$ according to definitions (5.4), (5.5). (Note: (5.9) requires that $x \sim w(\cdot)$ is independent of the randomness in R and z given x.) ■

Define

$$\rho(x_0|z) \equiv E\{\rho(x_0,x)|z\}. \tag{5.11}$$

The lemma says that

$$\log\{f_{x_0}(z)\} = \log\{f(z)\} + \log\{\rho(x_0|z)\} - \log\{\Pr\{\mathcal{R}\}\}, \tag{5.12}$$

yielding an extension of Tweedie's formula.

Theorem 5.2. *The conditional distribution $g_{x_0}(\eta|z)$ at the target value x_0 has cumulant generating function*

$$\lambda_{x_0}(z) = \lambda(z) + \log\{\rho(x_0|z)\} - \log\{\Pr\{\mathcal{R}\}\} \tag{5.13}$$

where $\lambda(z) = \log\{f(z)/f_0(z)\}$ as in (2.4).

Differentiating $\lambda_{x_0}(z)$ gives the conditional mean and variance of η,

$$\eta|z,x_0 \sim \left(l'(z) - l_0'(z) + \frac{d}{dz}\log\{\rho(x_0|z)\}, l''(z) - l_0''(z) + \frac{d^2}{dz^2}\log\{\rho(x_0|z)\}\right) \tag{5.14}$$

as in (2.7). For the normal translation family $z \sim \mathcal{N}(\mu,\sigma^2)$, formula (2.8) becomes

$$\mu|z,x_0 \sim \left\{z + \sigma^2\left[l'(z) + \frac{d}{dz}\log\{\rho(x_0|z)\}\right], \sigma^2\left(1 + \sigma^2\left[l''(z) + \frac{d^2}{dz^2}\log\{\rho(x_0|z)\}\right]\right)\right\}; \tag{5.15}$$

the estimate $E\{\mu|z,x_0\}$ is now the sum of the unbiased estimate z, an overall Bayes correction $\sigma^2 l'(z)$, and a further correction for relevance $\sigma^2(\log\{\rho(x_0|z)\})'$.

The relevance correction can be directly estimated in the same way as $\hat{l}'(z)$: first $\rho(x_0,x_i)$ is plotted versus z_i, then $\hat{\rho}(x_0|z)$ is estimated by a smooth regression and differentiated to give $(\log\{\hat{\rho}(x_0,x)\})'$. Figure 6 of Efron (2008b) shows a worked example of such calculations in the hypothesis-testing (rather than estimation) context of that paper.

In practice one might try various choices of the relevance function $\rho(x_0,x)$ to test their effects on $\hat{E}\{\mu|z,x_0\}$; perhaps $\exp\{-|x-x_0|/1000\}$ in the cnv example, or $I\{|x-x_0|/1000\}$ where I is the indicator function. This last choice could be handled by applying our original formulation (2.8) to only those z_i's having $|x_i - x_0| \le 1000$, as in Table 2. There are, however, efficiency advantages to using (5.15), particularly for narrow relevance definitions such as $|x_i - x_0| \le 100$; see Section 7 of Efron (2008b) for a related calculation.

Relevance need not be defined in terms of a single target value x_0. The covariates x in the brain scan example of Efron (2008b) are three-dimensional brain locations. For estimation within a region of interest A of the brain, say the hippocampus, we might set $\rho(x_0,x)$, for all $x_0 \in A$, to be some function of the distance of x from the nearest point in A.

A full Bayesian prior for the cnv data would presumably anticipate results like that for position 1755, automatically taking relevance into account in estimating μ_{1755}. Empirical Bayes applications expose the difficulties underlying this ideal. Some notion of *irrelevance* may become evident from the data, perhaps, for position 1755, the huge spike near position 4000 in Figure 6. The choice of $\rho(x_0, x)$, however, is more likely to be exploratory than principled: the best result, like that in Table 2, being that the choice is not crucial. A more positive interpretation of (5.15) is that Tweedie's formula $z + \sigma^2 l'(z)$ provides a general empirical Bayes correction for selection bias, which then can be fine-tuned using local relevance adjustments.

6 Tweedie's formula and false discovery rates

Tweedie's formula, and its application to selection bias, are connected with Benjamini and Hochberg's (1995) false discovery rate (Fdr) algorithm. This is not surprising: multiple testing procedures are designed to undo selection bias in assessing the individual significance levels of extreme observations. This section presents a Tweedie-like interpretation of the Benjamini–Hochberg Fdr statistic.

False discovery rates concern hypothesis testing rather than estimation. To this end, we add to model (2.1) the assumption that the prior density $g(\eta)$ includes an atom of probability π_0 at $\eta = 0$,

$$g(\eta) = \pi_0 I_0(\eta) + \pi_1 g_1(\eta) \qquad [\pi_1 = 1 - \pi_0] \tag{6.1}$$

where $I_0(\cdot)$ represents a delta function at $\eta = 0$, while $g_1(\cdot)$ is an "alternative" density of non-zero outcomes. Then the marginal density $f(z)$ (2.3) takes the form

$$f(z) = \pi_0 f_0(z) + \pi_1 f_1(z) \qquad \left[f_1(z) = \int_{\mathcal{Z}} f_\eta(z) g_1(\eta)\, d\eta \right]. \tag{6.2}$$

The *local false discovery rate* $\mathrm{fdr}(z)$ is defined to be the posterior probability of $\eta = 0$ given z, which by Bayes rule equals

$$\mathrm{fdr}(z) = \pi_0 f_0(z)/f(z). \tag{6.3}$$

Taking logs and differentiating yields

$$-\frac{d}{dz} \log\left(\mathrm{fdr}(z)\right) = l'(z) - l_0'(z) = E\{\eta | z\} \tag{6.4}$$

as in (2.6), (2.7), i.e., Tweedie's formula. Section 3 of Efron (2009) and Section 11.3 of Efron (2010b) pursue (6.4) further.

False discovery rates are more commonly discussed in terms of tail areas rather than densities. Let $F_0(z)$ and $F(z)$ be the right-sided cumulative distribution functions (or "survival functions") corresponding to $f_0(z)$ and $f(z)$. The right-sided *tail area Bayesian Fdr* given observation Z is defined to be

$$P(z) = \Pr\{\eta = 0 | Z \geq z\} = \pi_0 F_0(z)/F(z) \tag{6.5}$$

(with an analogous definition on the left). $F_0(z)$ equals the usual frequentist p-value $p_0(z) = \Pr_{F_0}\{Z \geq z\}$. In this notation, (6.5) becomes

$$P(z) = \frac{\pi_0}{F(z)}\ p_0(z). \tag{6.6}$$

Typically, $\pi_0/F(z)$ will be big in statistically interesting situations, illustrating the complaint, as in Berger and Sellke (1987), that p-values understate actual Bayesian null probabilities. This is particularly true in large-scale testing situations, where π_0 is often near 1.

The Benjamini–Hochberg Fdr control algorithm depends on an estimated version of $P(z)$,

$$\hat{P}(z) = \frac{\hat{\pi}_0}{\hat{F}(z)}\, p_0(z) \tag{6.7}$$

where $\hat{F}(z) = \#\{z_i \geq z\}/N$ is the usual empirical estimate of $F(z)$; π_0 can be estimated from the data, but more often $\hat{\pi}_0$ is set equal to 1. See for example Chapter 4 of Efron (2010b), where $\hat{P}(z)$ is considered as an empirical Bayes estimate of $P(z)$.

In the same spirit as (2.9), we can think of (6.7) as modifying the frequentist p-value $p_0(z)$ by an estimated Bayes correction factor $\hat{\pi}_0/\hat{F}(z)$, in order to obtain the empirical Bayes posterior probability $\hat{P}(z)$. Taking logarithms puts (6.7) into additive form,

$$-\log\left(\hat{P}(z_i)\right) = -\log\left(p_0(z_i)\right) + \log\left(\hat{F}(z_i)/\hat{\pi}_0\right) \tag{6.8}$$

for case i. (This is more general than Tweedie's formula in not requiring the structural models (2.1) and (6.1), rather only the specification of a null hypothsis density $f_0(z)$.)

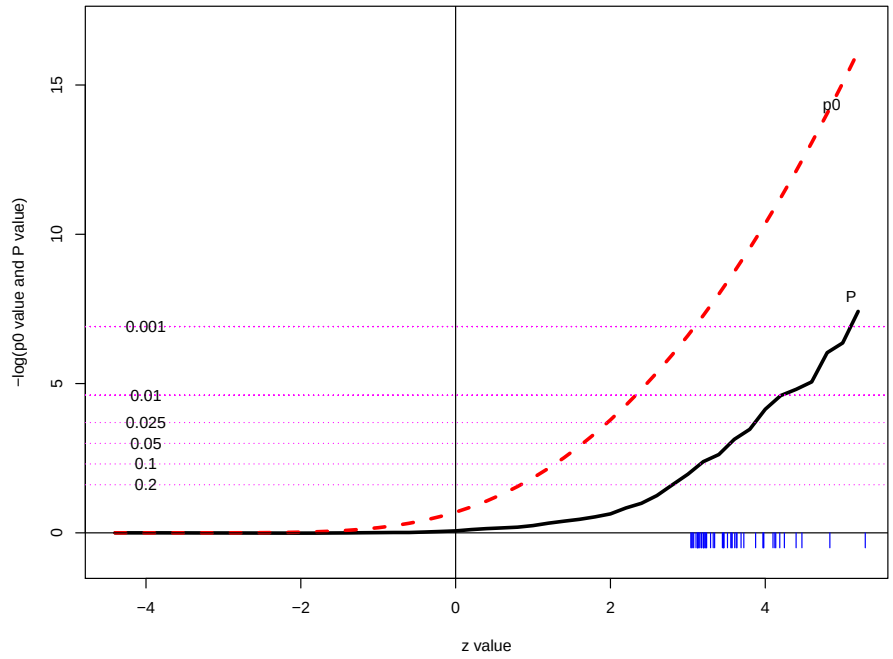

Figure 8: Frequentist p-value $p_0(z)$ (dashed curve) and the Benjamini–Hochberg empirical Bayes significance estimate $\hat{P}(z)$ (6.7) (solid curve), plotted on $-\log$ scale as in (6.8). Horizontal dashed lines indicate significant level on original scale. For prostate data, $N = 6033$ z-values from null hypothesis $f_0(z) = \mathcal{N}(0,1)$. Hash marks show the 49 z_i's exceeding 3.

Figure 8 applies (6.8), with $\hat{\pi}_0 = 1$, to the prostate data discusssed in Section 1 of Efron (2009). Here there are $N = 6033$ z_i's, with the null hypothesis $f_0(z)$ being a standard $\mathcal{N}(0,1)$ distribution. The histogram of the z_i's pictured in Figure 1 of Efron (2009) is a heavy-tailed

version of $f_0(z)$, with 49 cases exceeding $z = 3$, as indicated by the hash marks in Figure 8. The empirical Bayes correction is seen to be quite large: $z_i = 3$ for example has $p_0(z) = 0.00135$ but $\hat{P}(z) = 0.138$, one hundred times greater. A graph of Tweedie's estimate $z + \hat{l}'(z)$ for the prostate data (Efron, 2009, Fig. 2) is nearly zero for z between -2 and 2, emphasizing the fact that most of the 6033 cases are null in this example.

7 Remarks

This section presents some remarks expanding on points raised previously.

A. *Robbins' Poisson formula* Robbins (1956) derives the formula

$$E\{\mu|z\} = (z+1)f(z+1)/f(z) \tag{7.1}$$

for the Bayes posterior expectation of a Poisson variate observed to equal z. (The formula also appeared in earlier Robbins papers and in Good (1953), with an ackowledgement to A.M. Turing.) Taking logarithms gives a rough approximation for the expectation of $\eta = \log(\mu)$,

$$E\{\eta|z\} \doteq \log(z+1) + \log f(z+1) - \log f(z) \doteq \log(z+1/2) + l'(z) \tag{7.2}$$

as in (2.15).

B. *Skewness effects* Model (2.10) helps quantify the effects of skewness on Tweedie's formula. Suppose, for convenience, that $z = 0$ and $\sigma^2 = 1$ in (2.8), and define

$$I(c) = l'(0) + c\sqrt{1 + l''(0)}. \tag{7.3}$$

With $c = 1.96$, $I(c)$ would be the upper endpoint of an approximate 95% two-sided normal-theory posterior interval for $\mu|z$. By following through (2.12), (2.13), we can trace the change in $I(c)$ due to skewness. This can be done exactly, but for moderate values of the skewness γ, endpoint (7.3) maps, approximately, into

$$I_\gamma(c) = I(c) + \frac{\gamma}{2}\left(1 + I(c)^2\right). \tag{7.4}$$

If (7.3) gave the normal-model interval $(-1.96, 1.96)$, for example, and $\gamma = 0.20$, then (7.4) would yield the skewness-corrected interval $(-1.48, 2.44)$. This compares with $(-1.51, 2.58)$ obtained by following (2.12), (2.13) exactly.

C. *Correlation effects* The empirical Bayes estimation algorithm described in Section 3 does not require independence among the z_i values. Fitting methods like that in Figure 3 will still provide nearly unbiased estimates of $\hat{l}(z)$ for correlated z_i's, however, with increased variability compared to independence. In the language of Section 4, the empirical Bayes information per "other" observation is reduced by correlation. Efron (2010a) provides a quantitative assessment of correlation effects on estimation accuracy

D. *Total Bayes risk* Let $\mu^+(z) = z + l'(z)$ be the Bayes expectation in the normal model (2.8) with $\sigma^2 = 1$, and $\mathcal{R} = E_f\{(\mu - \mu^+)^2\}$ the total Bayes squared error risk. Then (2.8) gives

$$\begin{aligned}\mathcal{R} &= \int_{\mathcal{Z}} \left[1 + l''(z)\right] f(z)\, dz = \int_{\mathcal{Z}} \left[1 - l'(z)^2 + f''(z)/f(z)\right] f(z)\, dz \\ &= \int_{\mathcal{Z}} \left[1 - l'(z)^2\right] f(z)\, dz = 1 - E_f\left\{l'(z)^2\right\}\end{aligned} \tag{7.5}$$

under mild regularity conditions on $f(z)$. One might think of $l'(z)$ as the *Bayesian score function* in that its squared expectation determines the decrease below 1 of $\mathcal{R}$, given prior distribution g. Sharma (1973) and Brown (1971) discuss more general versions of (7.5).

***E**. Local estimation of the Bayes correction* The Bayes correction term $l'(z_0)$ can be estimated *locally*, for values of z near z_0, rather than globally as in (3.1). Let z_0 equal the bin center x_{k_0}, using the notation of (3.2)–(3.3), and define $K_0 = (k_1, k_2, \ldots, k_m)$ as the indices corresponding to a range of values $\boldsymbol{x}_0 = (x_{k_1}, x_{k_2}, \ldots, x_{k_m})$ near x_{k_0}, within which we are willing to assume a local linear Poisson model,

$$y_k \overset{\text{ind}}{\sim} \text{Poi}(\nu_k), \qquad \log(\nu_k) = \beta_0 + \beta_1 x_k. \tag{7.6}$$

Formula (4.4) is easy to calculate here, yielding

$$\text{Reg}(z_0) \doteq [N_0 \,\text{var}_0]^{-1}, \qquad \text{var}_0 = \sum_{K_0} \nu_k x_k^2/N_0 - \left(\sum_{K_0} \nu_k x_k/N_0\right)^2, \tag{7.7}$$

$N_0 = \sum_{K_0} \nu_k$. Local fitting gave roughly the same accuracy as the global fit of Section 3, with the disadvantage of requiring some effort in the choice of K_0.

***F**. Asymptotic regret calculation* The Poisson regression model (3.2)–(3.3) provides a well-known formula for the constant $c(z_0)$ in (4.4), $\text{Reg}(z_0) = c(z_0)/N$. Suppose the model describes the log marginal density $l(z)$ in terms of basis functions $B_j(z)$,

$$l(z) = \sum_{j=0}^{J} \beta_j B_j(z) \tag{7.8}$$

with $B_0(z) = 1$; $B_j(z) = z^j$ in (3.1), and a B-spline in Figure 3. Then

$$l'(z) = \sum_{j=0}^{J} \beta_j B'(z), \qquad B'_j(z) = dB_j(z)/dz. \tag{7.9}$$

We denote $l_k = l(x_k)$ for z at the bin center x_k, as in (3.3), and likewise l'_k, B_{j_k}, and B'_{j_k}.

Let $\boldsymbol{x}_k$ indicate the row vector

$$\boldsymbol{x}_k = (B_{0k}, B_{1k}, \ldots, B_{Jk}) \tag{7.10}$$

with $\boldsymbol{x}'_k = (B'_{0k}, B'_{1k}, \ldots, B'_{Jk})$, and $\boldsymbol{X}$ and $\boldsymbol{X}'$ the matrices having $\boldsymbol{x}_k$ and $\boldsymbol{x}'_k$ vectors, respectively, as rows. The MLE $\hat{\boldsymbol{\beta}}$ has approximate covariance matrix

$$\text{cov}\left(\hat{\boldsymbol{\beta}}\right) \doteq G^{-1}/N \qquad \text{where } G = \boldsymbol{X}^t \text{diag}(\boldsymbol{f}) \boldsymbol{X}, \tag{7.11}$$

$\text{diag}(\boldsymbol{f})$ the diagonal matrix with entries $\exp(l_k)$. But $l'_k = \boldsymbol{x}'_k \hat{\boldsymbol{\beta}}$, so $c_k = c(z_0)$ at $z_0 = x_k$ equals

$$c(x_k) = \boldsymbol{x}' G^{-1} \boldsymbol{x}'^t. \tag{7.12}$$

The heavy curve in Figure 5 was calculated using (4.5) and (7.12).

G. *Variable* σ^2 Tweedie's formula (1.4), $E\{\mu|z\} = z + \sigma^2 l'(z)$, assumes that σ^2 is constant in (1.2). However, σ^2 varies in the cnv example of Section 5, increasing from 5^2 to 7^2 as μ (i.e., k) goes from 20 to 60; see Figure 5 of Efron and Zhang (2011). The following theorem helps quantify the effect on the posterior distribution of μ given z.

Suppose (1.2) is modified to

$$\mu \sim g(\mu) \quad \text{and} \quad z|\mu \sim \mathcal{N}(\mu, \sigma_\mu^2) \tag{7.13}$$

where σ_μ^2 is a known function of μ. Let z_0 be the observed value of z, writing $\sigma_{z_0}^2 = \sigma_0^2$ for convenience, and denote the posterior density of μ given z under model (1.2), with σ^2 fixed at σ_0^2, as $g_0(\mu|z)$.

Theorem 7.1. *The ratio of* $g(\mu|z)$ *under model* (7.13) *to* $g_0(\mu|z)$ *equals*

$$\frac{g(\mu|z_0)}{g_0(\mu|z_0)} = c_0 \lambda_\mu \exp\left\{-\frac{1}{2}\left(\lambda_\mu^2 - 1\right)\Delta_\mu^2\right\} \tag{7.14}$$

where

$$\lambda_\mu = \sigma_0/\sigma_\mu \quad \textit{and} \quad \Delta_\mu = (\mu - z_0)/\sigma_0. \tag{7.15}$$

The constant c_0 *equals* $g(z_0|z_0)/g_0(z_0|z_0)$.

Proof. Let $\varphi(z;\mu,\sigma^2) = (2\pi\sigma^2)^{-1/2}\exp\{-\frac{1}{2}(\frac{z-\mu}{\sigma})^2\}$. Then Bayes rule gives

$$\frac{g(\mu|z_0)}{g(z_0|z_0)} = \frac{g(\mu)}{g(z_0)}\,\frac{\varphi(z_0;\mu,\sigma_\mu)}{\varphi(z_0;z_0,\sigma_0)} \tag{7.16}$$

and

$$\frac{g_0(\mu|z_0)}{g_0(z_0|z_0)} = \frac{g(\mu)}{g(z_0)}\,\frac{\varphi(z_0;\mu,\sigma_0)}{\varphi(z_0;\sigma_0,\sigma_0)}. \tag{7.17}$$

Dividing (7.16) by (7.17) verifies Theorem 7.1. ■

Figure 9 applies Theorem 7.1 to the Student-t situation discussed in Section 5 of Efron (2010a); we suppose that z has been obtained by normalizing a non-central t distribution having ν degrees of freedom and non-centrality parameter δ (not δ^2):

$$z = \Phi^{-1}\left(F_\nu(t)\right), \qquad t \sim t_\nu(\delta). \tag{7.18}$$

Here Φ and F_ν are the cdf's of a standard normal and a *central* Student-t distribution, so $z \sim \mathcal{N}(0,1)$ if $\delta = 0$. It is shown that (7.13) applies quite accurately, with σ_μ always less than 1 for $\delta \neq 0$. At $\nu = 20$ and $\delta = 5$, for instance, $\mu = 4.01$ and $\sigma_\mu = 0.71$.

It is supposed in Figure 9 that $\nu = 20$, and that we have observed $z = 3.0$, $l'(z) = -1$, and $l''(z) = 0$. The point estimate satisfying the stability relationship

$$\hat{\mu} = z + \sigma_{\hat{\mu}}^2 \cdot l'(z) \tag{7.19}$$

is calculated to be $\hat{\mu} = 2.20$, with $\sigma_{\hat{\mu}} = 0.89$. The dashed curve shows the estimated posterior density

$$\mu|z \sim \mathcal{N}\left(\hat{\mu}, \sigma_{\hat{\mu}}^2\right) \tag{7.20}$$

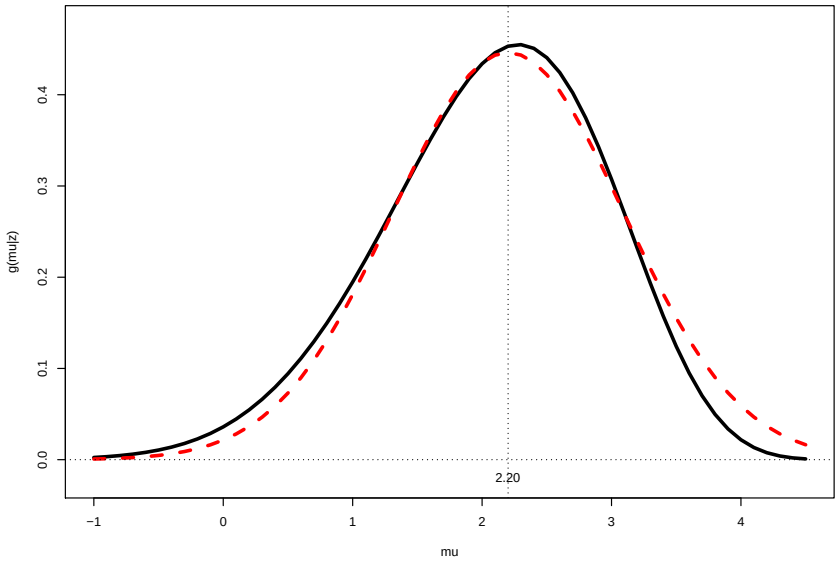

Figure 9: Application of Tweedie's formula to Student-t situation (7.18), having observed $z = 3$, $l'(z) = -1$, $l''(z) = 0$; stable estimate of $E\{\mu|z\}$ (7.19) is $\hat{\mu} = 2.20$, with $\sigma_{\hat{\mu}} = 0.89$. Dashed curve is estimate $g_0(\mu|z) \sim \mathcal{N}(\hat{\mu}, \sigma^2_{\hat{\mu}})$; solid curve is $g(\mu|z)$ from Theorem 7.1.

while the solid curve is the modification of (7.20) obtained from Theorem 7.1. In this case there are only modest differences.

Z-values are ubiquitous in statistical applications. They are generally well-approximated by model (7.13), as shown in Theorem 2 of Efron (2010a), justifying the use of Tweedie's formula in practice.

H. *Corrected expectations in the fdr model* In the false discovery rate model (6.1)–(6.3), let $E_1\{\eta|z\}$ indicate the posterior expectation of η given z and also given $\eta \neq 0$. Then a simple calculation using (2.7) shows that

$$\begin{aligned} E_1\{\eta|z\} &= E\{\eta|z\}/\left[1 - \text{fdr}(z)\right] \\ &= \left[l'(z) - l'_0(z)\right]/\left[1 - \text{fdr}(z)\right]. \end{aligned} \tag{7.21}$$

In the normal situation (2.8) we have

$$E_1\{\mu|z\} = \left[z + \sigma^2 l'(z)\right]/\left[1 - \text{fdr}(z)\right]. \tag{7.22}$$

The corresponding variance $\text{var}_1\{\mu|z\}$ is shown, in Section 6 of Efron (2009), to typically be smaller than $\text{var}\{\mu|z\}$. The normal approximation

$$\mu|z \doteq \mathcal{N}\left(E\{\mu|z\}, \text{var}\{\mu|z\}\right) \tag{7.23}$$

is usually inferior to a two-part approximation: $\mu|z$ equals 0 with probability $\text{fdr}(z)$, and otherwise is approximately $\mathcal{N}(E_1\{\mu|z\}, \text{var}\{\mu|z\})$.

I. *Empirical Bayes confidence intervals* In the James–Stein situation (3.7)–(3.9), the posterior distribution of μ having observed $z = z_0$ is

$$\mu_0|z_0 \sim \mathcal{N}(Cz, C) \qquad \text{where } C = 1 - 1/V. \tag{7.24}$$

We estimate C by $\hat{C} = 1 - (N-2)/\sum z_i^2$, so a natural approximation for the posterior distribution is

$$\mu_0|z_0 \dot{\sim} \mathcal{N}\left(\hat{C}z_0, \hat{C}\right). \tag{7.25}$$

This does not, however, take account of the estimation error in replacing C by $\hat{C}$. The variance of $\hat{C}$ equals $2/(V^2(N-4))$, leading to an improved version of (7.25),

$$\mu_0|z_0 \dot{\sim} \mathcal{N}\left(\hat{C}z_0, \hat{C} + \frac{2z_0^2}{V^2(N-4)}\right) \tag{7.26}$$

where the increased variance widens the posterior normal-theory confidence intervals. Morris (1983) derives (7.26) (with $N-4$ replaced by $N-2$) from a hierarchical Bayes formulation of the James–Stein model.

The added variance in (7.26) is the James–Stein regret (4.8). In the general context of Section 4, we can improve empirical Bayes confidence intervals by adding the approximate regret $c(z_0)/N$ (4.4) to estimated variances such as $1 + \hat{l}''(z_0)$. This is a frequentist alternative to full hierarchical Bayes modeling.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, **57** 289–300.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: irreconcilability of P values and evidence. *J. Amer. Statist. Assoc.*, **82** 112–139. With comments and a rejoinder by the authors, URL `http://links.jstor.org/sici?sici=0162-1459(198703)82:397<112:TAPNHT>2.0.CO;2-Z&origin=MSN`.

Brown, L. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Ann. Math. Statist.*, **42** 855–903.

Dawid, A. (1994). Selection paradoxes of Bayesian inference. In *Multivariate Analysis and Its Applications (Hong Kong, 1992)*, vol. 24 of *IMS Lecture Notes Monogr. Ser.* Inst. Math. Statist., Hayward, CA, 211–220.

Efron, B. (2008a). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.*, **23** 1–22.

Efron, B. (2008b). Simultaneous inference: When should hypothesis testing problems be combined? *Ann. Appl. Statist.*, **2** 197–223. `http://pubs.amstat.org/doi/pdf/10.1214/07-AOAS141`.

Efron, B. (2009). Empirical bayes estimates for large-scale prediction problems. *J. Amer. Statist. Assoc.*, **104** 1015–1028. `http://pubs.amstat.org/doi/pdf/10.1198/jasa.2009.tm08523`.

Efron, B. (2010a). Correlated z-values and the accuracy of large-scale statistical estimates. *J. Amer. Statist. Assoc.*, **105** 1042–1055. `http://pubs.amstat.org/doi/pdf/10.1198/jasa.2010.tm09129`.

Efron, B. (2010b). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, vol. I of *Institute of Mathematical Statistics Monographs*. Cambridge University Press, Cambridge.

Efron, B. and Zhang, N. R. (2011). False discovery rates and copy number variation. *Biometrika*, **98** 251–271.

Good, I. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, **40** 237–264.

Marshall, A. W. and Olkin, I. (2007). *Life Distributions*. Springer Series in Statistics, Springer, New York. Structure of nonparametric, semiparametric, and parametric families.

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *J. Amer. Statist. Assoc.*, **78** 47–65. With discussion.

Muralidharan, O. (2009). High dimensional exponential family estimation via empirical Bayes. Submitted, `http://stat.stanford.edu/~omkar/SSSubmission.pdf`.

Robbins, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I*. University of California Press, Berkeley and Los Angeles, 157–163.

Senn, S. (2008). A note concerning a selection "paradox" of Dawid's. *Amer. Statist.*, **62** 206–210.

Sharma, D. (1973). Asymptotic equivalence of two estimators for an exponential family. *Ann. Statist.*, **1** 973–980.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, **9** 1135–1151.

Sun, L. and Bull, S. (2005). Reduction of selection bias in genomewide studies by resampling. *Genet. Epidemiol.*, **28** 352–367.

van Houwelingen, J. C. and Stijnen, T. (1983). Monotone empirical Bayes estimators for the continuous one-parameter exponential family. *Statist. Neerlandica*, **37** 29–43. `http://dx.doi.org/10.1111/j.1467-9574.1983.tb00796.x`.

Zhang, C.-H. (1997). Empirical Bayes and compound estimation of normal means. *Statist. Sinica*, **7** 181–193. Empirical Bayes, sequential analysis and related topics in statistics and probability (New Brunswick, NJ, 1995).

Zhong, H. and Prentice, R. L. (2008). Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics*, **9** 621–634.

Zhong, H. and Prentice, R. L. (2010). Correcting "winner's curse" in odds ratios from genomewide association findings for major complex human diseases. *Genet. Epidemiol.*, **34** 78–91.

Zollner, S. and Pritchard, J. K. (2007). Overcoming the winner's curse: Estimating penetrance parameters from case-control data. *Amer. J. Hum. Genet.*, **80** 605–615.