Empirical Bayes: Concepts and Methods

Bradley Efron

Abstract

Empirical Bayes methods originated in the work of Robbins, Good and Toulmin, Stein and others in the 1950s. This overview of the main ideas focuses on some empirical Bayes success stories: the missing species problem, James–Stein shrinkage estimators, and Bayesian false discovery rates. Different modeling strategies are reviewed, as well as the dual frequentist/Bayesian nature of the methods. Technical issues are kept to a minimum, at the expense of avoiding most theoretical developments. The final section discusses the problem of *relevance*– how to decide which "other" observations apply to inference about a particular case of interest.

1 Introduction

Robbins (1956) coined the name "empirical Bayes" for his method of dealing with the following kind of inferential situation: an unknown probability density $g(\theta)$ ("density" here including the possibility of discrete atoms) has produced a random sample of realizations $\theta_1, \theta_2, \ldots, \theta_N$; the θ_i are unobserved but each one has yielded an observed random variable x_i according to a known family of densities $f(x_i \mid \theta_i)$:

$$\theta_i \sim g(\theta) \quad \text{and} \quad x_i \sim f(x_i \mid \theta_i) \tag{1.1}$$

independently for $i = 1, 2, \ldots, N$. From the observed sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ we wish to estimate the parameters $\theta_1, \theta_2, \ldots, \theta_N$.

Robbins' monumental insight was that, in addition to x_i, the other observations x_j, $j \neq i$, could assist in the estimation of θ_i. Table 1 concerns one of his examples. An automobile insurance company has tabulated the number of claims made by each of its 9461 customers during the past year, with x_i the number of claims made by customer i. The table shows that 7840 of the customers made $x = 0$ claims, 1317 made $x = 1$ claim, etc., up to the one customer, perhaps not the most dependable driver, who made

Table 1: Counts y_x of the number of insurance claims made in a single year by 9461 policy holders; Robbins' formula (1.3) estimates the number of claims in the following year. For instance, a customer who had made $x = 2$ claims was expected to make 0.527 claims the next year. *Gamma MLE* is from Fisher's gamma model, Section 2.

Claims x	0	1	2	3	4	5	6	7
Counts y_x	7840	1317	239	42	14	4	4	1
Formula (1.3)	0.168	0.363	0.527	1.33	1.43	6.00	1.25	–
Gamma MLE	0.165	0.376	0.669	0.94	1.16	1.32	1.45	1.55

seven claims. Now the insurance company wants to set premiums for the succeeding insurance year. Of course it would be helpful to know how many claims each customer could be expected to make.

Let y_x be the number of customers making x claims in the past year, $y_0 = 7840$, $y_1 = 1317$, etc. Robbins supposed that $f(x_i \mid \theta_i)$ in (1.1) was Poisson,

$$f(x_i \mid \theta_i) = e^{-\theta_i}\theta_i^{x_i}/x_i!, \tag{1.2}$$

$x_i = 0, 1, 2, \ldots$, θ_i being the Poisson expectation for customer i. (Bad drivers have high θ_i values.) Without any assumptions about the prior density $g(\theta)$ in (1.1), he showed that the Bayes posterior expected value for θ_i given x_i, $E\{\theta_i \mid x_i\}$, could be estimated by

$$\widehat{E}\{\theta_i \mid x_i\} = (x_i + 1)y_{x_i+1}/y_{x_i}. \tag{1.3}$$

So, for instance, a driver who had had no accidents could be expected to have $\widehat{E}\{\theta_i \mid 0\} = y_1/y_0 = 0.168$ accidents in the coming year. Robbins' derivation of formula (1.3) appears at the end of this section.

Robbins' compact little formula has a certain magical quality. It raises some intriguing ideas that we'll be dealing with in what follows:

- The insurance company's data set, the line of counts in Table 1, by itself provides the information for estimating $E\{\theta_i \mid x_i\}$, a Bayesian quantity, without appeal to prior knowledge. In other words, the estimate $\widehat{E}\{\theta_i \mid x_i\}$ is purely frequentistic (doesn't involve prior assumptions), justifying the name "empirical Bayes". Frequentist or not, somehow we've produced a Bayesian result without any Bayesians. The idea that a large data set could imply its own prior distribution was revolutionary.

- The estimate $\widehat{E}\{\theta_i \mid x_i\}$ for driver i depends on his or her past experience x_i, but also on the experience of the other drivers. A driver with a clean record, $x_i = 0$, has estimate $\widehat{E}\{\theta_i \mid x_i = 0\} = y_1/y_0$, partly depending on less-clean drivers who had had one accident each. There is a tacit assumption being made on the relevance of others' experiences to any one individual's insurance premium; see Section 6.

- Robbins' formula is nonparametric in the sense that it doesn't depend on any parametric form for $g(\theta)$. In fact, $g(\theta)$ is never estimated in the proof of (1.3). Robbins' derivation, at the end of this section, finesses the role of $g(\theta)$.

- Formula (1.3) goes astray at the right side of Table 1, where the counts y_x get small. The more stable "Gamma MLE" estimates begin by parametrically modeling $g(\theta)$ as a gamma distribution, as explained in Section 2 and discussed in Section 5.

- The insurance company doesn't care if formula (1.3) is individually accurate–only if it is close enough overall to give a good picture of forthcoming claims. Any one driver, though, might hope for individual applicability. This distinction becomes more salient if, say, we are talking about individual versus overall prognosis of a serious medical condition.

This paper is not intended to be a systematic review of the empirical Bayes literature. The basic idea of empirically estimating part of a Bayesian procedure can take on forms other than (1.1). A great amount of effort, starting with Robbins' original papers, has gone into verifying the asymptotic accuracy of formulas like (1.3), but none of that is part of what follows. Here I will focus on some notable empirical Bayes success stories and the conceptual and methodological issues they raise.

Derivation of Robbins' formula If $\theta \sim g(\theta)$, where $g(\theta)$ is the prior density, and $x \mid \theta \sim f(x \mid \theta)$, then according to Bayes rule, the posterior density of θ given x is

$$g(\theta \mid x) = g(\theta) f(x \mid \theta)/f(x), \tag{1.4}$$

where $f(x)$ is the marginal density of x,

$$f(x) = \int_\Theta f(x \mid \theta) g(\theta)\, d\theta, \tag{1.5}$$

with Θ the set of possible θ values.

For the Poisson density $f(x \mid \theta) = \theta^x e^{-\theta}/x!$ we get

$$\begin{aligned} f(x) &= \int_0^\infty (e^{-\theta}\theta^x/x!)g(\theta)\ d\theta \\ \text{and} \qquad g(\theta \mid x) &= g(\theta)\frac{e^{-\theta}\theta^x/x!}{f(x)}. \end{aligned} \tag{1.6}$$

This yields the posterior expectation

$$\begin{aligned} E\{\theta \mid x\} = \int_0^\infty \theta g(\theta \mid x)\ d\theta &= \frac{\int_0^\infty \left(e^{-\theta}\theta^{x+1}/x!\right) g(\theta)\ d\theta}{f(x)} \\ &= (x+1)\frac{\int_0^\infty \left(e^{-\theta}\theta^{x+1}/(x+1)!\right) g(\theta)\ d\theta}{f(x)} \\ &= (x+1)f(x+1)/f(x). \end{aligned} \tag{1.7}$$

In the empirical Bayes situation (1.1) we don't know $f(x)$ or $f(x+1)$ but they are easy to estimate. The expected value of y_x/N is $f(x)$. Substituting y_x/N for $f(x)$ and y_{x+1}/N for $f(x+1)$ in the bottom line of (1.7) gives Robbins' formula (1.3). We will see this same tactic — expressing a posterior quantity of interest directly in terms of the marginal density $f(x)$ — show up again in Section 2 through Section 4, as a way of avoiding estimating the prior $g(\theta)$. Section 5 discusses methods that *do* estimate $g(\theta)$.

2 The missing species problem

Robbins named empirical Bayes and made the concept explicit but an earlier paper can be recognized as a clear empirical Bayes triumph: Fisher, Corbet, and Williams (1943). Corbet, a leading naturalist, had been trapping butterflies in Malaysia (then Malaya) for two years. Table 2 shows his trapping data, in the same format as Table 1. Let

$$y_x = \text{number of species trapped } x \text{ times} \tag{2.1}$$

during the two years: $y_1 = 118$ species were so rare they had been trapped just one time each, $y_2 = 74$ trapped twice each, up to the three species trapped 24 times each. (We will take Table 2 as comprising all the data though in fact Corbet also had counts for the more common, less interesting species.)

Table 2: Butterfly data; number y_x of species seen x times each in two years of trapping; 118 species trapped just once, 74 trapped twice each, etc.

x	1	2	3	4	5	6	7	8	9	10	11	12
y_x	118	74	44	24	29	22	20	19	20	15	12	14
x	13	14	15	16	17	18	19	20	21	22	23	24
y_x	6	12	6	9	9	6	10	10	11	5	3	3

There is a crucial difference between Table 2 and Table 1: in the latter case there is no $x = 0$ entry. These are the "missing species". Corbet asked Fisher a seemingly unanswerable missing species question: if he continued trapping for one additional year, how many new species could he expect to see? Fisher answered Corbet using what we would now call a parametric empirical Bayes model. We will return to Fisher's solution, but first present a nonparametric method due to Good and Toulmin (1956).

Let i index the species, say

$$i = 1, 2, \ldots, S, \tag{2.2}$$

S being the total number of species both observed and missing. The basic assumption of both the parametric and nonparametric analyses is that species i is being trapped according to a Poisson process with intensity parameter θ_i.

Let x_i be the number of times species i is trapped in the original one unit of time (the unit being two years for Corbet), and $x_i(t)$ the number of times trapped in a hypothetical future period of t units. The Poisson process assumption says that

$$\Pr\{x_i = x\} = e^{-\theta_i}\theta_i^x/x! \tag{2.3}$$

and

$$\Pr\{x_i = 0 \text{ and } x_i(t) > 0\} = e^{-\theta_i}\left[1 - e^{-\theta_i t}\right]. \tag{2.4}$$

This last is the probability that species i is *not* seen in the initial trapping period but *is* seen in the future period, i.e., it is one of Corbet's hoped-for "new" species.

The prior density $g(\theta)$ can be thought of as an empirical density putting probability $1/S$ in each of the S parameters θ_i. We can write $E\{t\}$, the

expected number of new species seen in an additional trapping period of t units, as

$$E\{t\} = S \int_0^\infty e^{-\theta}[1 - e^{-\theta t}]g(\theta)\ d\theta \tag{2.5}$$

(ignoring the discreteness in $g(\theta)$). Expanding $[1 - e^{-\theta t}]$ in a Taylor series gives

$$E\{t\} = S \int_0^\infty e^{-\theta} \left[\theta t - \frac{(\theta t)^2}{2} + \frac{(\theta t)^3}{3!} - \ldots\right] g(\theta)\ d\theta. \tag{2.6}$$

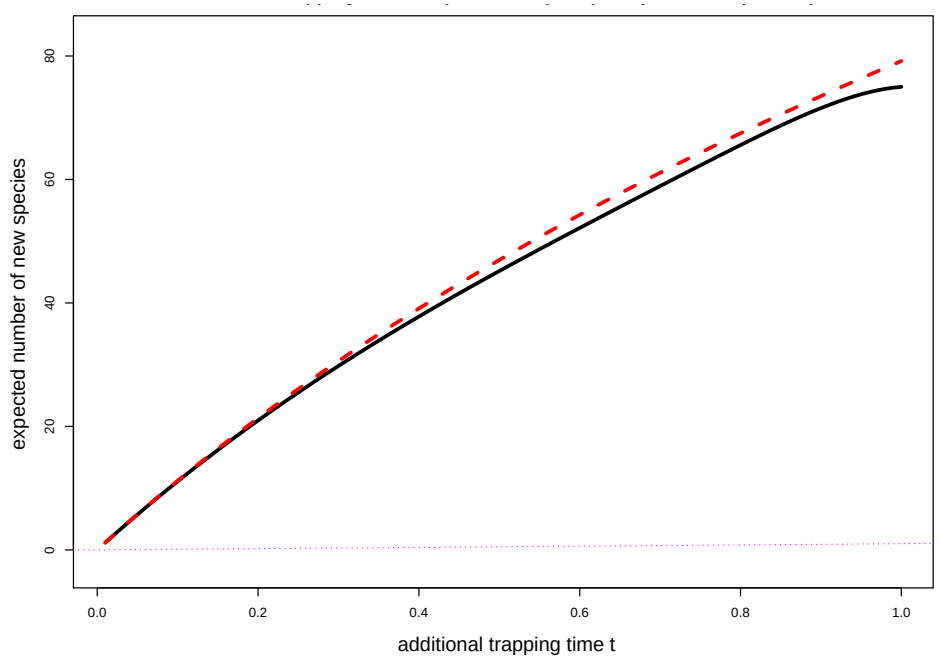

Figure 1: Expected number of new butterfly species seen in t units additional trapping time; nonparametric (solid), parametric (dashed).

Summing (2.3) over the S species gives an expression for e_x, the expected value of y_x,

$$e_x = E\{y_x\} = S \int_0^\infty \left[e^{-\theta}\theta^x/x!\right] g(\theta)\ d\theta. \tag{2.7}$$

Comparing (2.6) and (2.7), we can write

$$E\{t\} = e_1 t - e_2 t^2 + e_3 t^3 - \ldots; \tag{2.8}$$

e_x is unbiasedly estimated by the count y_x, leading to Good and Toulmin's intriguing estimator[1]

$$\widehat{E}\{t\} = y_1 t - y_2 t^2 + y_3 t^3 - \ldots. \tag{2.9}$$

Corbet asked Fisher about one more year of trapping; that is 1/2 of a 2-year time unit, giving an estimated number of newly trapped butterfly species $\widehat{E}\{0.5\}$ equaling

$$\widehat{E}\{0.5\} = 118 \cdot 0.5 - 74 \cdot 0.5^2 + 44 \cdot 0.5^3 - \cdots = 45.2. \tag{2.10}$$

For two more trapping years, $t = 1$, formula (2.9) gives $\widehat{E}\{1\} = 75$. The heavy curve in Figure 1 graphs $\widehat{E}\{t\}$ for t between 0 and 1. Formula (2.9) diverges for t greater than 1, a defect considered next.

Fisher's answer to Corbet's missing species question began with the assumption that the prior density $g(\theta)$ for the Poisson process parameters θ_i had a scaled gamma form,

$$g_{\alpha,\beta}(\theta) = \frac{\theta^{\alpha-1} e^{-\theta/\beta}}{\beta^\alpha \Gamma(\alpha)} \qquad (\theta > 0). \tag{2.11}$$

That is, $\theta \sim \beta G_\alpha$ for G_α a standard gamma variate with shape parameter α. In this case, changing parameters from (α, β) to (α, γ) with $\gamma = \beta/(1+\beta)$, the marginal density $f(x)$ (1.5) is

$$f_{\alpha,\gamma}(x) = c_{\alpha,\gamma} \frac{\Gamma(x+\alpha)}{x!\Gamma(\alpha)} \gamma^x (1-\gamma)^\alpha \qquad [\gamma = \beta/(1+\beta)], \tag{2.12}$$

proportional to a negative binomial density, namely the probability of getting x tails before α heads, flipping a coin with probability γ of tails. The proportionality constant $c_{\alpha,\gamma}$ is necessary in the missing species problem because the data is truncated; for Corbet's data, $c_{\alpha,\gamma}$ equals one over the sum for x from 1 to 24 of the untruncated negative binomial density.

Thinking of the counts y_x in Table 2 as a multinomial sample, maximization of the log likelihood

$$\sum_{x=1}^{24} y_x \log\left(f_{\alpha,\gamma}(x)\right) \tag{2.13}$$

[1]Evidently, the missing species problem also arose in World War II code-breaking efforts: Good credited his fellow code-breaker Alan Turing for some of the ideas here. The argument leading to (2.9) — substituting observed counts for their expectations — is similar to that for Robbins' formula, though the authors worked separately.

gave maximum likelihood estimates

$$\hat{\alpha} = 0.104 \quad \text{and} \quad \hat{\gamma} = 0.989, \qquad \text{so } \hat{\beta} = \frac{\hat{\gamma}}{(1-\hat{\gamma})} = 89.79. \tag{2.14}$$

Fisher then used the prior density $g_{\hat{\alpha},\hat{\beta}}(\theta)$ to estimate $E\{t\}$ by substitution in (2.5).

There is no particular reason to think that the gamma prior (2.11) applies to Corbet's data, but in this case it nicely agrees with Good and Toulmin's nonparametric solution, as seen in Figure 1. For $t = 1/2$, one year of additional trapping, it gives $\widehat{E}\{1/2\} = 46.95$ compared with 45.2 before.

Fisher's gamma model provides a neat formula for $\widehat{E}\{t\}$,

$$\widehat{E}\{t\} = \frac{y_1\left[1 - \frac{1}{(1+\hat{\gamma}t)^{\alpha}}\right]}{\hat{\gamma}\hat{\alpha}}. \tag{2.15}$$

This gives reasonable-looking predictions for t greater than 1, for instance

$$\widehat{E}\{2\} = 123, \qquad \widehat{E}\{4\} = 176. \tag{2.16}$$

"Reasonable" does *not* necessarily mean convincing. The MLE value $\hat{\alpha} = 0.104$ endows $g_{\hat{\alpha},\hat{\beta}}(\theta)$ with an enormous spike near zero, representing very rare species. As t increases, more and more of the estimate $\widehat{E}\{t\}$ depends on the exact shape of the spike. Letting t go to infinity in (2.15) gives

$$\widehat{E}\{\infty\} = 11422, \tag{2.17}$$

an estimate of the total number of species S: not something one wants to bet on. The problem of very rare species becomes more acute in the next example, where the species are words in Shakespeare's vocabulary.

Table 3 provides a statistical view of the Shakespearean canon.[2] Altogether the canon comprises 884647 words, of which there are 31534 *distinct* words. (Henceforth, "words" will mean "distinct words".) The upper left entry shows that 14376 words were so rare they appeared just once each in the canon; 4343 appeared twice each, 2292 three times each, up to 5 words that appeared 100 times each; 846 words appeared more than 100 times each, but these play no role in what follows.

We can ask Corbet's question for Shakespeare: suppose one found some previously unknown Shakespeare amounting to proportion t of the canon,

[2]These numbers are based on the Concordance in Spevack (1968); differences in what is considered "Shakespeare" have altered the counts a little since then.

Table 3: Shakespeare's word counts; 14376 distinct words appeared once each in the canon, 4343 distinct words twice each, etc.; the canon contains 31534 distinct words, with 884647 words in total, counting repeats.

	1	2	3	4	5	6	7	8	9	10
0+	14376	4343	2292	1463	1043	837	638	519	430	364
10+	305	259	242	223	187	181	179	130	127	128
20+	104	105	99	112	93	74	83	76	72	63
30+	73	47	56	59	53	45	34	49	45	52
40+	49	41	30	35	37	21	41	30	28	19
50+	25	19	28	27	31	19	19	22	23	14
60+	30	19	21	18	15	10	15	14	11	16
70+	13	12	10	16	18	11	8	15	12	7
80+	13	12	11	8	10	11	7	12	9	8
90+	4	7	6	7	10	10	15	7	7	5

that is, $884647 \cdot t$ total words, counting repeats. (This would be the equivalent of a new trapping period of length t.) How many distinct new words, not among the original 31534, would we expect to find? Letting t go to infinity, the question becomes, "How many words did Shakespeare know but not use?"

For $t = 1$, i.e., a find of new Shakespeare equal in length to the canon, Good and Toulmin's nonparametric estimator (2.9) gives

$$\widehat{E}\{1\} = 11430. \tag{2.18}$$

Fisher's Gamma model is again in good agreement, giving

$$\widehat{E}\{1\} = 11483.^{3} \tag{2.19}$$

The maximum likelihood estimates for the gamma prior $g_{\alpha,\beta}(\theta)$ were

$$\hat{\alpha} = -0.3954 \quad \text{and} \quad \hat{\beta} = 104.26 \tag{2.20}$$

($\hat{\gamma} = 0.9905$). Now the prior $g_{\hat{\alpha},\hat{\beta}}(\theta)$ is improper, the spike near zero being infinite–this is allowable since we don't have to deal with the missing species

[3] All this might seem like pure speculation. However, cross-validation tests were run: one of the plays was removed from the canon, Table 3 recomputed for the remainder, and predictions $\widehat{E}\{t\}$ made for the omitted play. Accuracy of both parametric and nonparametric predictions was impressive.

count y_0. As $t \to \infty$, Fisher's estimate $\widehat{E}\{t\}$ goes to infinity as more and more hypothetical cases are pulled out of the infinite spike.

One sees that there can't be a way to pin down $E\{\infty\}$, the number of words Shakespeare knew but didn't use. However, Efron and Thisted (1976) used linear programming methods to put a nonparametric lower bound on $E\{\infty\}$,

$$E\{\infty\} > 35554. \tag{2.21}$$

Adding in the 31534 distinct words in the canon, this implies a Shakespearean vocabulary exceeding 67000 words.[4]

In November of 1985, noted Shakespearean scholar Gary Taylor found a short love poem of 429 words, "Shall I fly?", which he attributed to Shakespeare. This proved immediately controversial, no new Shakespearean work having been discovered since the 1600s (and the poem not being very good).

Table 4: Counts y_x and their estimate $\widehat{E}_x$ (2.9) for the poem "Shall I fly?"; 9 distinct words in the poem never appeared in the canon ($x = 0$), with (2.9) predicting 6.97.

x	0	1	2	3	4	5	6	7	8	9
y_x	9	7	5	4	4	2	4	0	2	3
$\widehat{E}_x$	6.97	4.21	3.33	2.84	2.53	2.43	2.16	2.01	1.87	1.76

Thisted and Efron (1987) analyzed the poem's rare word counts using Good and Toulmin's approach. Here the new trapping period t is quite small, $t = 429/884647$, a little less than 0.0005. Formula (2.9) predicts $\widehat{E}\{t\} = 6.97$ while nine distinct words in the poem had actually never appeared in the canon. Good and Toulmin's method also provides predictions for the number of words in the poem that appeared once each in the canon, twice each, etc., as shown in Table 4. The predictions are consistently lower than the observations, but other than that they seem reasonably agreeable.

Thisted and Efron concluded that Shakespearean authorship could not be ruled out on the basis of rare word counts. "Shall I fly?" has never been accepted into the canon, but Taylor and others have gone on to use statistical methods in a variety of the Bard's attribution questions such as a possible collaboration with Marlowe on Parts I, II, and III of *Henry VI*.

[4] "Words" here means any distinct spelling, and "persons" a different word than "person". Contemporaries such as Marlowe and Jonson appear to have used larger vocabularies than Shakespeare.

The authoritative *New Oxford Shakespeare* does indeed list Marlowe as a collaborator, to the fury of scholars less swayed by statistics.

Table 5: A sequence analysis combining data from 10 mice gave a library of 237,863 distinct clones, of which 142,500 occurred only once each, 28,524 occurred twice each, etc. If another 10 mice were sequenced, Good and Toulmin's nonparametric method estimates seeing about 123,000 previously unseen distinct clones.

x	1	2	3	4	5	6	7	8	9	10
y_x	142500	28524	13296	7699	5161	3709	2785	2217	1822	1510
x	11	12	13	14	15	16	17	18	19	20
y_x	1235	1151	1067	942	874	785	724	655	659	542

The missing species problem shows up in biogenetical investigations. Table 5 concerns a sequence analysis of mouse DNA. Ten mice had their genomes sequenced and split into "clones", i.e., short fragments. Some of the clones were common but 142,500 others were so rare they only showed up once in the combined pool of ten mice; 28,524 appeared twice each, 13,296 three times each, etc. Application of Good and Toulmin's formula (2.9) showed that the observed library of 287,863 clones was by no means complete: a new set of ten mice could be expected to yield $\widehat{E}\{1\} = 122,900$ previously unseen clones.

The differences between Good and Toulmin's and Fisher's approaches go further than nonparametric versus parametric. Fisher's method actually estimates the prior $g(\theta)$ while Good and Toulmin's doesn't. With $\hat{g}(\theta)$ in hand, the statistician can assess Bayesian quantities other than $E\{t\}$, for instance $\Pr\{\theta \le 1 \mid x\}$. More on the two empirical Bayes approaches, *f-modeling* and *g-modeling*, will be seen in Section 6.

3 The James–Stein estimator

Early developments of empirical Bayes methodology featured Poisson distributions for observations x_i in model (1.1). The James–Stein estimator (James and Stein, 1961) focused instead on normally distributed observations,

$$x_i \overset{\text{ind}}{\sim} \mathcal{N}(\theta_i, \sigma_0^2) \qquad (i = 1, 2, \ldots, N), \tag{3.1}$$

though the empirical Bayes connection took a while to develop.

Faced with situation (3.1), and without any prior knowledge of how differently the unknown parameters θ_i might arrange themselves, the obvious estimators are

$$\hat{\theta}_i^{\text{MLE}} = x_i \qquad \text{for } i = 1, 2, \ldots, N, \tag{3.2}$$

the "MLE" superscript noting that these are maximum likelihood estimates. Their expected total squared error is

$$E\left\{\sum_{i=1}^{N}\left(\hat{\theta}_i^{\text{MLE}} - \theta_i\right)^2\right\} = N\sigma_0^2. \tag{3.3}$$

(We will take σ_0^2 as known.)

James and Stein's 1961 paper, building on Stein's earlier work, proved the following result:

Theorem 1. *Let*

$$\hat{\theta}_i^{\text{JS}} = \bar{x} + \left(1 - \frac{(N-3)\sigma_0^2}{S}\right)(x_i - \bar{x}) \qquad \left(S = \sum_{i=1}^{N}(x_i - \bar{x})^2\right). \tag{3.4}$$

Then, assuming model (3.1) *with* $N \geq 4$,

$$E\left\{\sum_{i=1}^{N}\left(\hat{\theta}_i^{\text{JS}} - \theta_i\right)^2\right\} < N\sigma_0^2 \tag{3.5}$$

for all possible choices of the θ_i*'s.*

The James–Stein theorem has a good claim to being the most striking and disruptive statistics result of the post-war era: striking in its simplicity and disruptive of conventional wisdom. A century of experience with normal theory OLS regression models, all of which employ estimates like (3.2), seemed suddenly to be on shaky ground. One could do better, at least if "better" is defined by expected square error loss. (That's a crucial caveat, as we'll see in Section 6.)

If the improvement offered by the James–Stein estimates were infinitesimal, only theorists would be interested. In favorable situations the gains can, in fact, be enormous. Table 6 (Efron and Morris, 1972) concerns one such situation. The batting averages of 18 major league baseball players were recorded over the 1970 season.[5] Column 1 of the table shows their batting averages after 45 attempts, about one-tenth of the full season. Player i's

[5] *Batting average* is the number of successful hits per attempt, or at-bat. Clemente, for instance, hit successfully in 18 of his first 45 at-bats, for an average of $18/45 = 0.400$ ("batting .400"). Not knowing anything about baseball is an advantage for getting the full force of this example.

Table 6: Batting averages of 18 baseball players. *Column 1*, observed batting average after 45 attempts. *Column 2*, James–Stein estimate (3.4), $\sigma_0^2 = 0.066$. *Column 3*, average for remainder of the 1970 season. *Column 4*, number of attempts in remainder of the season.

	MLE	JS	final avg	# final AB
Clemente	.400	.294	.346	367
F. Robinson	.378	.289	.298	426
F. Howard	.356	.284	.276	521
Johnstone	.333	.279	.222	275
Berry	.311	.275	.273	418
Spencer	.311	.275	.270	466
Kessinger	.289	.270	.263	586
L. Alvarado	.267	.265	.210	138
Santo	.244	.261	.269	510
Swoboda	.244	.261	.230	200
Unser	.222	.256	.264	277
Williams	.222	.256	.256	270
Scott	.222	.256	.303	435
Petrocelli	.222	.256	.264	538
E. Rodriguez	.222	.256	.226	186
Campaneris	.200	.251	.285	558
Munson	.178	.246	.316	408
Alvis	.156	.242	.200	70

observed average x_i can be thought of as the maximum likelihood estimate $\hat{\theta}_i^{\mathrm{MLE}}$ of θ_i, his true batting average. The grand average of the 18 observed averages was

$$\bar{x} = \sum_{i=1}^{18} x_i/18 = 0.265. \tag{3.6}$$

Column 2 gives the James–Stein estimates $\hat{\theta}_i^{\mathrm{JS}}$ (3.4), taking

$$\sigma_0 = \left[\frac{\bar{x}\,(1-\bar{x})}{45}\right]^{1/2} = 0.066, \tag{3.7}$$

the binomial standard deviation based on 45 coin flips with probability of success $\bar{x}$. We don't know the true averages but as a surrogate, Column 3 lists the players' final averages over the remaining nine-tenths of the 1970 regular season, $\theta_i^{\mathrm{final}}$.

The prediction errors for $\hat{\theta}^{\rm MLE}$ and $\hat{\theta}^{\rm JS}$ were

$$\sum_{i=1}^{18}\left(\hat{\theta}_i^{\rm MLE}-\theta_i^{\rm final}\right)^2=0.0754 \quad \text{and} \quad \sum_{i=1}^{18}\left(\hat{\theta}_i^{\rm JS}-\theta_i^{\rm final}\right)^2=0.0213. \tag{3.8}$$

In this case using the James–Stein rule reduces prediction error by a factor of more than 3. Nobody cares much about 50-year-old baseball statistics, but if we were predicting cancer cure rates at 18 hospitals, three times less error might be life-saving.

James and Stein's proof of Theorem 1 is an elegant exercise in frequentist inference. Efron and Morris (1972) motivated formula (3.4) from a less elegant empirical Bayes viewpoint. They envisioned a "normal-normal" version of setup (1.1),

$$\theta_i\sim\mathcal{N}(M,A) \quad \text{and} \quad x_i\sim\mathcal{N}(\theta_i,\sigma_0^2), \tag{3.9}$$

independently for $i=1,2,\ldots,N$. That is, the prior density $g(\theta)$ is assumed to be normal with mean M and variance A, both M and A unknown to the statistician.

The Bayes estimate of θ_i under model (3.9) is

$$\hat{\theta}_i^{\rm Bayes}=M+B(x_i-M) \qquad \left[B=A/(A+\sigma_0^2)\right]. \tag{3.10}$$

The marginal distribution of the x_i under (3.9) is

$$x_i\overset{\rm ind}{\sim}\mathcal{N}(M,V) \qquad (V=A+\sigma_0^2), \tag{3.11}$$

so that

$$\bar{x}\sim\mathcal{N}(M,V/N) \qquad \text{independent of } S\sim V\chi^2_{N-1}, \tag{3.12}$$

with S as in (3.4). Parameters M and B in (3.10) are estimated unbiasedly by

$$\widehat{M}=\bar{x} \quad \text{and} \quad \widehat{B}=\left[1-\frac{(N-3)\sigma_0^2}{S}\right], \tag{3.13}$$

the last following from

$$E\left\{\left(\chi^2_{N-1}\right)^{-1}\right\}=(N-3)^{-1}.$$

Using (3.13) in (3.4), we can rewrite the James–Stein estimator as

$$\hat{\theta}_i^{\rm JS}=\widehat{M}+\widehat{B}\left(x_i-\widehat{M}\right). \tag{3.14}$$

The term "empirical Bayes" for the James–Stein rule seems entirely appropriate, $\hat{\theta}_i^{\rm JS}$ being $\hat{\theta}_i^{\rm Bayes}$ (3.10), with the prior hyperparameters B and M frequentistically estimated from the observed sample $x_1, x_2, \ldots, x_n$.

The James–Stein theorem is a wonderfully exact result: the simple rule (3.4) *always* bests the MLE, with no exceptions or further assumptions. Exact results exert a powerful attraction in statistics, a sprawling and usually inexact science. The immediate history of Theorem 1 involved searching for other estimators that dominated maximum likelihood, particularly admissible ones (which (3.4) isn't; see Strawderman, 1971).

And yet it is an *inexact* aspect of (3.4) that perhaps has had the greatest long-term effect: that *shrinkage* is a powerful tool for the simultaneous estimation of multiple parameters. The baseball example violates the assumptions of Theorem 1 — the batting averages are binomial rather than normal — but the savings accrue nonetheless.

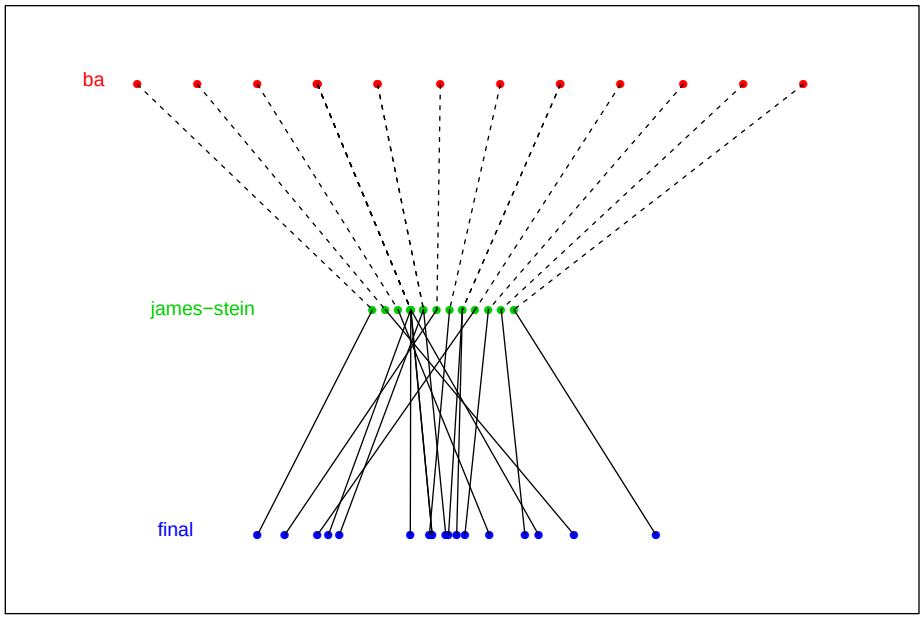

Figure 2: Batting averages of 18 baseball players. *Top*, observed batting averages after 45 attempts. *Middle*, James–Stein estimates. *Bottom*, average for remainder of the 1970 season.

Shrinkage is quite extreme for the baseball players, each estimate $\hat{\theta}_i^{\rm JS}$ being shrunk 78% $(= 1 - \widehat{B})$ of the way from x_i to the grand mean $\bar{x}$. This

is illustrated in Figure 2, where we can see that the James–Stein rule is an *over-shrinker*, a favorable tactic for total squared error risk it inherits from Bayes rule (3.10). The tactic works because the unbiased estimates x_i are, as a group, overdispersed,

$$\begin{aligned} &\frac{E\left\{\sum_{i=1}^{N}(x_i-\bar{x})^2\right\}}{N-1} = A+\sigma_0^2, \\ \text{compared to} \qquad &\frac{E\left\{\sum_{i=1}^{N}\left(\theta_i-\bar{\theta}\right)^2\right\}}{N-1} = A. \end{aligned} \tag{3.15}$$

Regularization, the big brother of shrinkage, is used in modern estimation algorithms such as the lasso. Empirical Bayes methods tend toward strong regularization.

A good question: what if the normal prior $\theta_i \sim \mathcal{N}(M, A)$ in (3.9) isn't a good match to the problem at hand? Is there a version of the James–Stein rule applicable more generally? The answer is yes. Efficient empirical Bayes methods, supported by modern computational capabilities, do the job (though without the frequentist guarantees of the James–Stein theorem). An example, artificial but informative, follows next.

Figure 3 shows the "two towers" example: $N = 1500$ x_i values have been observed,

$$x_i \sim \mathcal{N}(\theta_i, 1) \qquad (i = 1, 2, \ldots, N = 1500), \tag{3.16}$$

where 500 of the θ_i's were uniformly selected in the interval $[-1.7, -0.7]$ and 1000 uniformly selected in $[0.7, 2.7]$. The red solid histogram shows the two towers of θ_i values, while the 1500 x_i values from (3.16) appear in the dashed black histogram.

The statistician wishes to estimate the θ_i's using some estimation rule $e(x)$,

$$\hat{\theta}_i = e(x_i) \qquad (i = 1, 2, \ldots, N), \tag{3.17}$$

with loss function the expected average squared error (EASE),

$$\text{EASE} = E\left\{\sum_{i=1}^{N}\left(\hat{\theta}_i - \theta_i\right)^2 / N\right\}, \tag{3.18}$$

the expectation being over model (3.16). The MLE rule $e(x_i) = x_i$ has

$$\text{EASE}_{\text{MLE}} = 1. \tag{3.19}$$

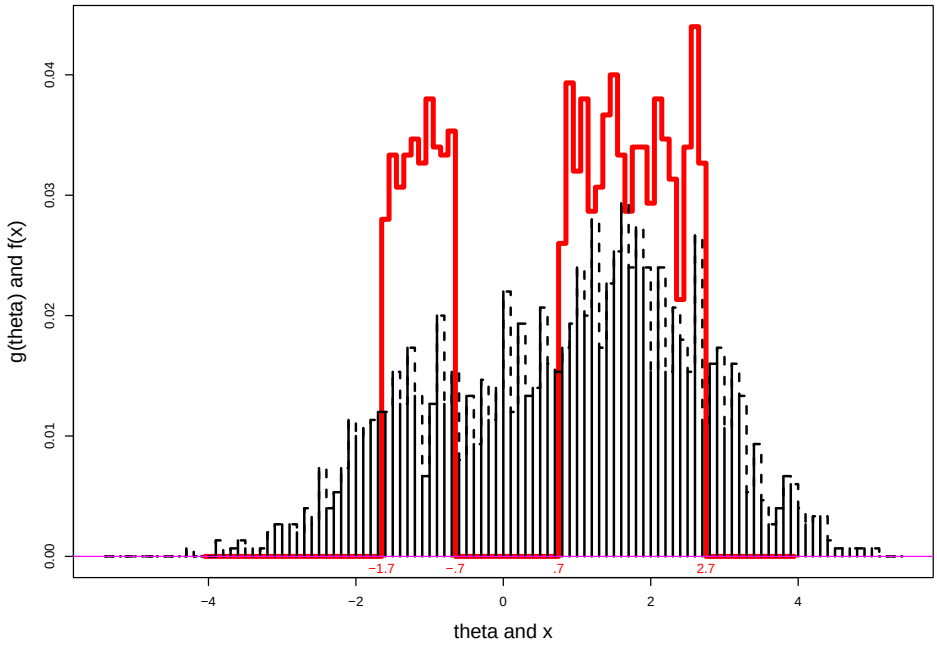

Figure 3: Two towers example. *Solid red histogram*, $N = 1500$ parameters θ_i. *Dashed black histogram*, observations $x_i \sim \mathcal{N}(\theta_i, 1)$.

Suppose though a friendly Oracle has told us the order statistic of the θ_i's

$$\theta_{(1)} \le \theta_{(2)} \le \cdots \le \theta_{(N)},$$

without letting on which $\theta_{(j)}$ goes with which θ_i. (Equivalently, the Oracle tells us the histogram of the θ_i's. Oracle calculations, as in Jiang and Zhang, 2009 and Brown and Greenshtein, 2009, go back to Robbins' original papers.) Let $\bar{g}(\theta)$ denote the discrete density putting probability $1/N$ on each $\theta_{(j)}$. The Oracle's information allows us to compute the Bayes posterior expectation of θ_i given x_i assuming prior $\bar{g}(\theta)$,

$$e_{\bar{g}}(x) = \frac{\sum_{j=1}^{N} \theta_{(j)} \varphi(x - \theta_{(j)})}{\sum_{j=1}^{N} \varphi(x - \theta_{(j)})}, \tag{3.20}$$

with $\varphi(x)$ the standard normal density $\exp(-x^2/2)/\sqrt{2\pi}$.

For the two tower prior in Figure 3, numerical calculations show that rule (3.20) has

$$\text{EASE}_{\bar{g}} = 0.563, \tag{3.21}$$

more than 40% less than EASE_{MLE}. There aren't any Oracles of course, but this is where modern empirical Bayes comes in, allowing us to do nearly as well as (3.21) using a fully data-based estimator. This amounts to extending the James–Stein rule to non-normal priors $g(\theta)$, though without Theorem 1's guaranteed superiority.

The Bayesian model underlying our empirical Bayes estimates is

$$\theta \sim g(\cdot) \quad \text{and} \quad x \mid \theta \sim \mathcal{N}(\theta, 1). \tag{3.22}$$

Given x, the conditional distribution has mean and variance denoted as

$$\theta \mid x \sim (e_g(x), v_g(x)) \,. \tag{3.23}$$

Let $f(x)$ indicate the marginal density of x (1.5),

$$f(x) = \int_{\mathcal{T}} \varphi(x-\theta) g(\theta)\, d\theta, \tag{3.24}$$

where $\mathcal{T}$ is the domain of θ. *Tweedie's formulas* provide convenient expressions for $e_g(x)$ and $v_g(x)$ in terms of $l(x) = \log f(x)$,

$$e_g(x) = x + \dot{l}(x) \quad \text{and} \quad v_g(x) = 1 + \ddot{l}(x), \tag{3.25}$$

$\dot{l}(x) = dl(x)/dx$ and $\ddot{l}(x) = d^2 l(x)/dx^2$; see Efron (2010).

If $\hat{\theta} = e(x)$ is some estimation rule for θ in model (3.22), its overall squared error risk $R(g, e)$ is

$$R(g, e) = E\left\{\left(\hat{\theta} - \theta\right)^2\right\}, \tag{3.26}$$

the theoretical version of EASE (3.18). The Bayes conditional expectation $e_g(x)$ minimizes (3.26), the Bayes risk equaling

$$\begin{aligned} R(g, e_g) = \int_{-\infty}^{\infty} v_g(s) f(x)\, dx &= \int_{-\infty}^{\infty} \left(1 + \ddot{l}(x)\right) f(x)\, dx \\ &= 1 - \int_{-\infty}^{\infty} \dot{l}(x)^2 f(x)\, dx, \end{aligned} \tag{3.27}$$

the last following from the general result $E\{\dot{l}^2\} = -E\{\ddot{l}\}$. Using $\hat{\theta} = e(x)$ not equal to $e_g(x)$ must increase $R(g, e)$, the *regret* being

$$R(g, e) - R(g, e_g) = \int_{-\infty}^{\infty} \left(e(x) - e_g(x)\right)^2 f(x)\, dx. \tag{3.28}$$

Section 2 of Efron (2019) verifies (3.27)–(3.28).

Figure 3 suggests an empirical Bayes estimation approach:

1. Estimate the marginal density $f(x)$ by a smooth curve $\hat{f}(x)$ drawn through the bar tops of the black histogram for $x_1, x_2, \ldots, x_{1500}$.

2. Estimate e_g by
$$\hat{e}(x) = x + \frac{d}{dx} \log \hat{f}(x), \tag{3.29}$$
as in (3.25). Section 5 discusses a convenient way to fit $\hat{f}(x)$ based on Poisson regression to the histogram bin counts.

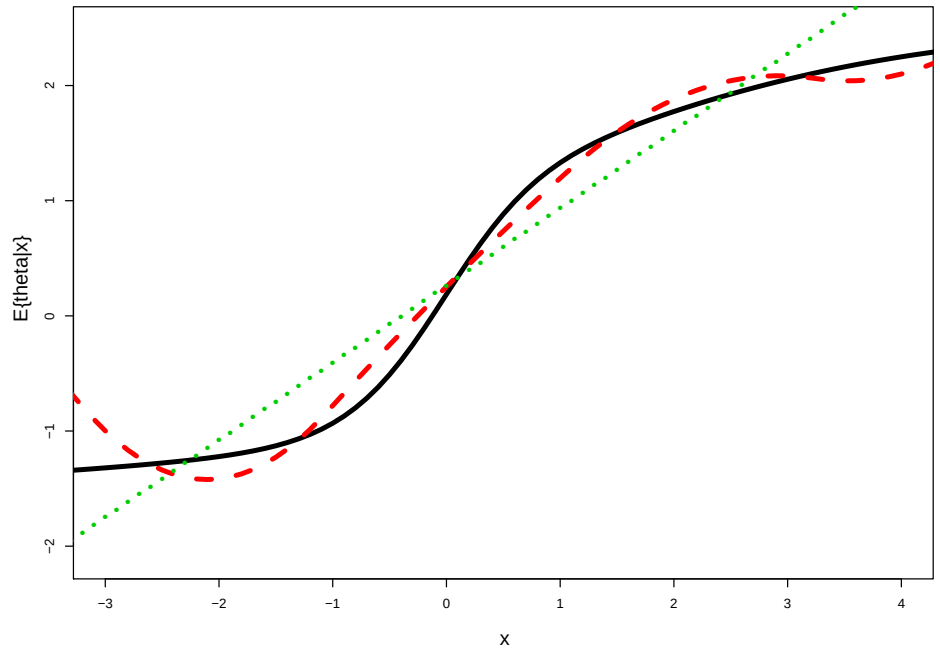

Figure 4: Posterior expectations $E\{\theta \mid x\}$: Oracle (black solid); estimated (red dashed) from a natural spline fit with 5 degrees of freedom (i.e., with 6 free parameters counting the intercept); James–Stein (green dotted).

The solid black curve in Figure 4 shows $e_g(x)$ (3.20), the Bayes posterior expectation based on the Oracle's empirical prior $\bar{g}(\theta)$. $R(\bar{g}, e_{\bar{g}}) = 0.563$, the value of $\text{EASE}_{\bar{g}}$ given previously. (It isn't difficult to show that this is the minimum possible value of EASE for any rule of the form $\hat{\theta} = e(x)$.) The red dashed curve is $\hat{e}(x)$, an empirical Bayes estimate (3.29), with $\hat{f}(x)$ obtained from a natural spline fit, five degrees of freedom, to the histogram bin counts of the 1500 x_i's in Figure 3.

The empirical Bayes rule $\hat{e}(x)$ does a reasonably good job of approximating $e_{\bar{g}}$, missing badly only in the lower 1% tail of the x_i values. Its empirical Bayes regret (3.28) was

$$\int_{-\infty}^{\infty} \left(\hat{e}(x) - e_{\bar{g}}(x)\right)^2 \bar{f}(x)\, dx = 0.026, \tag{3.30}$$

using the Oracle marginal density

$$\bar{f}(x) = \sum_{1}^{N} \frac{\varphi(x - \theta_{(i)})}{N}.$$

One can do better, but the estimated risk $0.563 + 0.026 = 0.589$ is still more than 40% less than the MLE risk 1.

The James-Stein rule (3.14) is the straight line graphed in green dots for Figure 4, and has empirical Bayes regret 0.117. The two towers prior is emphatically non-normal so it isn't surprising to see James–Stein erring here.

4 False discovery rates

The false discovery rate story is strikingly similar to that of the James–Stein estimator: a frequentist theorem that provides surprising and appealing data analysis possibilities, disruptive of prevailing wisdom, is later reinterpreted in empirical Bayes terms. The disruptive theorem in this case is the FDR control algorithm of Benjamini and Hochberg (1995), summarized next.

Suppose we have N independent testing situations, each of which provides a p-value p_i against its null hypothesis H_{0i}. Let $r(\boldsymbol{p})$ be some rule that inputs the vector $\boldsymbol{p}$ of N p-values, and outputs a decision for or against each H_{0i}, $i = 1, 2, \ldots, N$. The rule is observed to reject H_{0i} for R of the N tests. The *false discovery proportion* is then

$$\text{Fdp} = R_0/R, \tag{4.1}$$

where R_0 is the number of the R rejections where H_{0i} was actually true, i.e., the number of false discoveries. (R_0 and Fdp are not observable statistics.) The *false discovery rate* of rule $r(\cdot)$ is defined to be the expected value of Fdp,

$$\text{Fdr} = E\{\text{Fdp}\}, \tag{4.2}$$

with Fdp set equal to zero if $R = 0$.

The Benjamini–Hochberg algorithm is a rule $r_q(\boldsymbol{p})$ guaranteed to have Fdr $\leq q$ where q is some small pre-chosen constant like 0.1. The algorithm begins by ordering the p-values,

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(N)}. \tag{4.3}$$

Let $i_{\max}$ be the largest index i for which

$$p_{(i)} \leq (i/N)q; \tag{4.4}$$

rule r_q rejects $H_{0(i)}$ for $i \leq i_{\max}$, and accepts $H_{0(i)}$ for $i > i_{\max}$.

Theorem 2. *The rule r_q has*

$$\text{Fdr} = \hat{\pi}_0 q \qquad \textit{where } \hat{\pi}_0 = N_0/N, \tag{4.5}$$

N_0 the number of true null hypotheses.

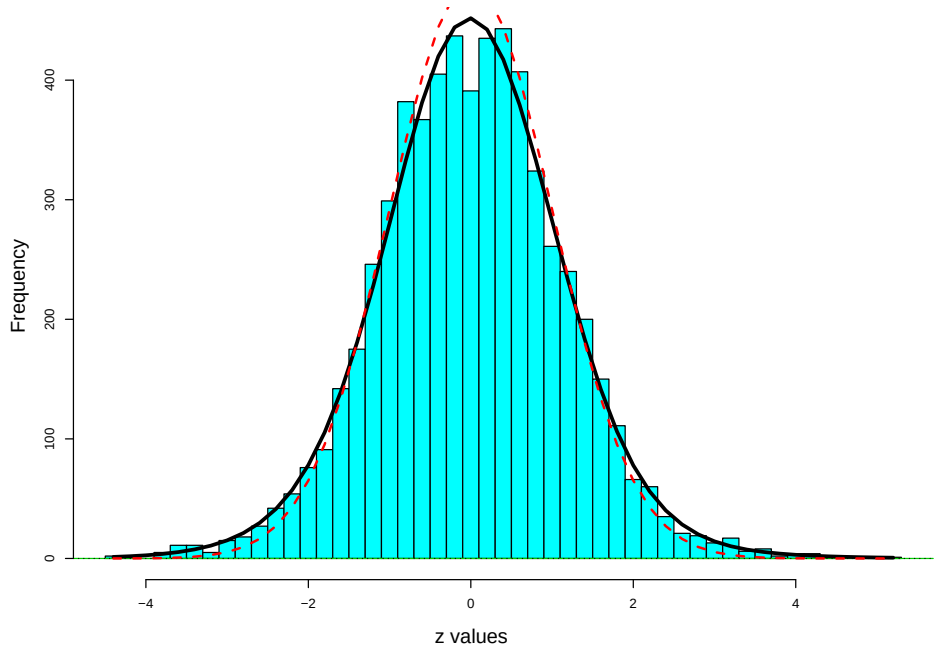

Figure 5: Frequencies for 6033 z-values in prostate microarray study; heavy curve for natural spline fit with 5 degrees of freedom; light dashed curve counts proportional to $\mathcal{N}(0,1)$ fit.

N_0 is unobservable but in the kind of "fishing expedition" where mass testing is often used, N_0/N is near 1. In any case, (4.5) is usually taken to say that r_q controls Fdr at level q.

As an example, the histogram in Figure 5 shows z-values from a microarray study involving 50 prostate cancer patients and 52 healthy controls. Genetic activities for 6033 genes were compared between cancer and control subjects, giving a z-value z_i for each of the $N = 6033$ genes, the null hypothesis being

$$H_{0i} : z_i \sim \mathcal{N}(0, 1), \qquad i = 1, 2, \ldots, N. \tag{4.6}$$

If all the H_{0i} were true the histogram would follow a $\mathcal{N}(0, 1)$ density, the dashed curve, which is seen to be too high in the center and too low in the tails. (Section 2.1 of Efron, 2010 gives a fuller description of the prostate cancer data.)

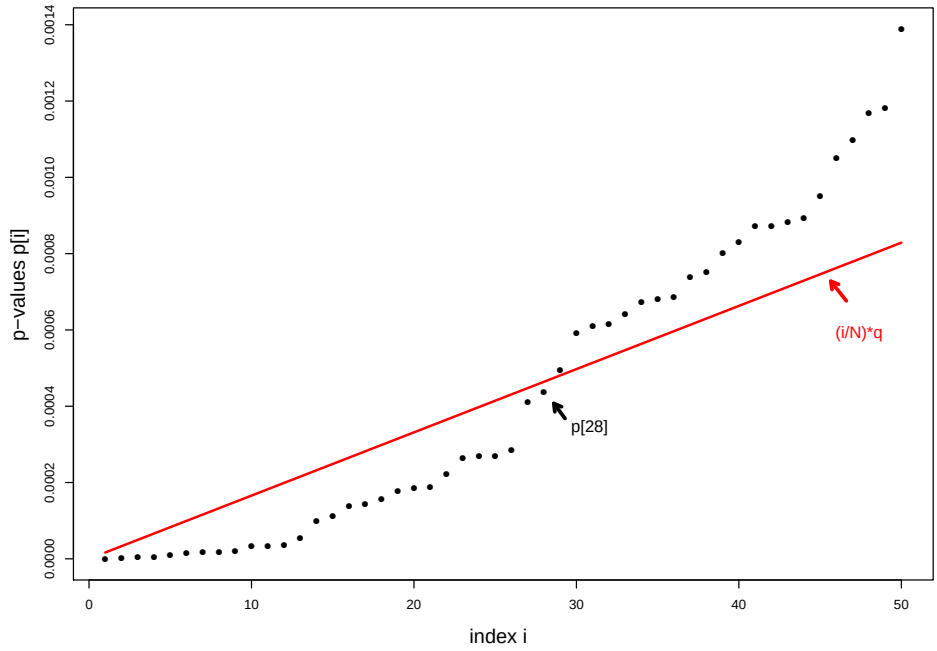

Figure 6: Benjamini–Hochberg rule r_q with $q = 0.1$ applied to right side of prostate data; rejects largest 28 z-values $z_i > 3.33$.

Figure 6 illustrates the Benjamini–Hochberg rule r_q, $q = 0.1$, applied to

the right side of the histogram; that is, using right-sided p-values

$$p_i = 1 - \Phi(z_i), \tag{4.7}$$

Φ the standard normal cdf. The ordered values $p_{(i)}$ are plotted versus index i. They cross the control line $(i/N)\cdot q$ just after $i = 28$, so r_q rejects $H_{0(i)}$ for $i = 1, 2, \ldots, 28$ and accepts for $i \geq 29$. The rejection threshold corresponds to $z_i > 3.33$.[6]

As with the James–Stein estimator, the Benjamini–Hochberg rule (4.4) can be motivated by a simple Bayesian model. We suppose that each z-value is randomly selected to be either "null" with prior probability π_0 or "non-null" with probability $\pi_1 = 1 - \pi_0$; nulls follow density $f_0(x)$ — the $\mathcal{N}(0,1)$ density $\varphi(z)$ in (4.6) — while non-nulls follow $f_1(z)$,

$$\pi_0 \longrightarrow f_0 \longrightarrow z; \qquad \pi_1 \longrightarrow f_1 \longrightarrow z. \tag{4.8}$$

Let $F_0(z)$ and $F_1(z)$ be the right-sided[7] cdfs,

$$F_0(z) = \int_z^\infty f_0(x)\ dx \quad \text{and} \quad F_1(z) = \int_z^\infty f_1(x)\ dx \tag{4.9}$$

and

$$F(z) = \pi_0 F_0(z) + \pi_1 F_1(z); \tag{4.10}$$

define $\mathrm{Fdr}(z)$ as the posterior probability of nullness given an observed z-value exceeding z,

$$\mathrm{Fdr}(z) = \Pr\{\text{null} \mid \text{observed } z\text{-value } \geq z\} = \pi_0 F_0(z)/F(z) \tag{4.11}$$

according to Bayes rule. Even if $F_0(z)$ is known (as is usually assumed in hypothesis testing situations) and π_0 is taken to be nearly 1, still the marginal cdf $F(z)$ will most often be unknown. However, there is an obvious nonparametric estimate of $F(z)$, the empirical cdf of the N z-values,

$$\widehat{F}(z) = \#\{z_i \geq z\}/N. \tag{4.12}$$

Plugging this into (4.11) gives the *empirical Bayes Fdr estimate*

$$\widehat{\mathrm{Fdr}}(z) = \hat{\pi}_0 F_0(z)/\widehat{F}(z), \tag{4.13}$$

[6]The familiar Bonferroni bound $p_i \leq 0.05/N$ rejects for $z_i \geq 4.40$, only the top 4 genes making the cutoff. The popularity of FDR has much to do with its relative liberality.

[7]Right-sided here to agree with the example in Figure 6 where we rejected for large positive values of z.

where $\hat{\pi}_0$ could be 1 in practice. We might decide to reject H_{0i} if the ratio $\widehat{\text{Fdr}}(z_i)/\hat{\pi}_0$ of posterior- to prior-nullness is smaller than some threshold q. In terms of the ordered values $z_{(i)}$, we reject H_{0i} if

$$F_0(z_{(i)})/\widehat{F}(z_{(i)}) \le q. \tag{4.14}$$

But $F_0(z_{(i)})$ is the p-value $p_{(i)}$, and $i/n = \widehat{F}(z_{(i)})$. Looking at (4.4), we see that the empirical Bayes rejection rule (4.14) is the same as the Benjamini–Hochberg rule r_q!

From a Bayesian point of view we should prefer the *local false discovery rate*

$$\text{fdr}(z) = \Pr\{\text{null} \mid z\} = \pi_0 f_0(x)/f(z), \tag{4.15}$$

to the tail area rate fdr(z) (4.11). Estimating fdr(z) requires a parametric estimate of the marginal density,

$$f(z) = \pi_0 f_0(z) + \pi_1 f_1(z),$$

as contrasted with the nonparametric $\widehat{F}(z)$ (4.12).

Lindsey's method (Lindsey, 1974) offers a convenient way to construct such estimates. The method begins by assuming an exponential family form for the density $f(z)$,

$$\log f_\beta(z) = \beta^\top s(z) - \psi(\beta), \tag{4.16}$$

where $s(z)$ is a vector of sufficient statistics; for example,

$$s(z) = (z, z^2, z^3, z^4, z^5) \tag{4.17}$$

would allow $\log f_\beta(z)$ to be any fifth-degree polynomial; $\psi(\beta)$ is whatever constant is needed to make $f_\beta(z)$ integrate to 1.

It looks like special software would be needed to calculate the MLE for β, but Lindsey's method finesses the problem by discretization. The histogram in Figure 5 has $K = 49$ bins, with bin k having, say centerpoint x_k and count y_k. Count y_k has expected value

$$\mu_k(\beta) = \exp\{\beta_0 + \beta^\top s_k\}, \tag{4.18}$$

$s_k = s(x_k)$, where β_0 is a proportionality constant chosen to make

$$\sum_1^N \mu_k(\beta) = N.$$

We assume that the counts y_k are independently Poisson with expectation (4.18). This is a multiparameter exponential family, for which standard GLM software gives the MLE $\hat{\beta}$, nearly identical to the MLE in family (4.16).

Model (4.16) was applied to the $N = 6033$ z-values of Figure 5, taking $s(z)$ in (4.16) to be a fifth-degree natural spline rather than polynomial (4.17). It gave MLE density $\hat{f}(z)$ and estimated local false discovery rate

$$\widehat{\text{fdr}}(z) = f_0(z)/\hat{f}(z), \tag{4.19}$$

with $f_0(z) = f(z)$ and $\pi_0 = 1$. This is shown in the left panel of Figure 7.[8]

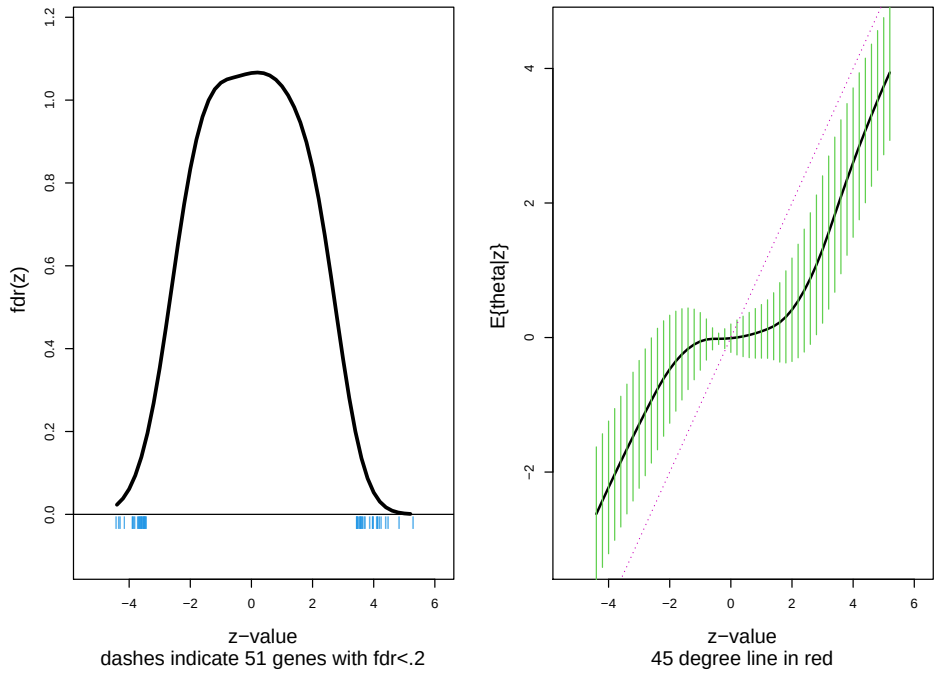

Figure 7: *Left panel*, estimated local false discovery rate fdr(z) for prostate cancer study (ns fit/5 df). *Right panel*, expected effect size $E\{\theta \mid z\}$; bars are +/− posterior standard deviation.

Fifty-one of the genes had $\text{fdr}(z_i) \leq 0.2$, a conventional threshold for non-nullness. How non-null were they? Suppose

$$z_i \sim \mathcal{N}(\theta_i, 1), \qquad i = 1, 2, \ldots, N, \tag{4.20}$$

[8]The fact that $\widehat{\text{fdr}}(z)$ slightly exceeds 1 near $z = 0$ demonstrates that π_0 is actually a little less than 1. This permits an estimate of π_0, ignored here.

θ_i the *effect size* for gene i. As in (3.22)–(3.25),

$$\begin{aligned} &\theta \mid z \dot{\sim} \left(\hat{e}(z), \hat{v}(z)\right), \\ \text{with} \qquad &\hat{e}(z) = z + \frac{d}{dz}\log \hat{f}(z) \quad \text{and} \quad \hat{v}(z) = 1 + \frac{d^2}{dz^2}\log \hat{f}(z). \end{aligned} \tag{4.21}$$

The right panel of Figure 7 shows $\hat{e}(z) \pm \hat{v}(z)^{1/2}$. At $z = 4$ for instance, the distribution of θ given z is estimated to have posterior expectation 2.60 and standard deviation 1.11.

The precise nature of Theorem 2 elicited a substantial literature devoted to exact frequentist control of FDR. Empirical Bayes methods are less exact but, as with the James–Stein rule, they open up a wider range of possibilities such as the local false discovery rate. An important difference is that Theorem 2 requires independence among the observations z_i while empirical Bayes expectation results like $\hat{e}(z)$ (4.21) do not.

5 f-modeling and g-modeling

Our four empirical Bayes success stories — Robbins' formula (1.3), the missing species estimate (2.9), the James–Stein rule (3.14), and false discovery rates (4.12) — all share one important characteristic: they avoid estimating the prior density $g(\theta)$ in the basic empirical Bayes setup (1.1),

$$g(\theta) \longrightarrow \theta_i \quad \text{and} \quad f(x \mid \theta_i) \longrightarrow x_i, \tag{5.1}$$

$i = 1, 2, \ldots, N$. Instead they focus on directly estimating the marginal density $f(x)$ from the observations

$$x_i \sim f(x), \qquad i = 1, 2, \ldots, N. \tag{5.2}$$

The reason is simple enough: estimating $g(\theta)$ can be hard work. In the classic situation where $f(x_i \mid \theta_i)$ is the normal kernel $\varphi(x_i - \theta_i)$, finding $g(\theta)$ is the famously difficult *deconvolution* problem. Estimating $g(\theta)$ in the general context (5.1) can be called the *empirical Bayes deconvolution* problem, as in Efron (2016).

And yet there are good reasons one might want an estimate of the prior $g(\theta)$, involving questions that can't be answered directly in terms of the marginal density $f(x)$. In the prostate cancer effect-size model $z_i \sim \mathcal{N}(\theta_i, 1)$ (4.16), what is $\Pr_g\{\theta_i \geq 2\}$ say, or $\Pr_g\{\theta_i \geq 2 \mid z_i = 3\}$?

The term *g-modeling* refers to methods that *do* begin with direct estimation of $g(\theta)$, as opposed to *f-modeling* which focuses on $f(x)$. Fisher's

parametric gamma prior for the missing species problem (2.11) was the initial g-modeling example. Nonparametric g-modeling, in which $g(\theta)$ is discretely represented by a small number of point masses, goes back to Robbins' original publication (Kiefer and Wolfowitz, 1956; Laird, 1978), and has enjoyed renewed interest in an age of computational advances (Koenker and Mizera, 2014). Here we will only discuss recent developments in parametric g-modeling (Efron, 2016).

The mechanics of g-modeling are easiest to explain in a discrete setting where θ can take on m possible values, say $\theta \in \mathcal{T}$,

$$\mathcal{T} = \left\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(m)}\right\}, \tag{5.3}$$

and x only n possible values, $x \in \mathcal{X}$,

$$\mathcal{X} = \left\{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\right\}. \tag{5.4}$$

Now $g(\theta)$ can be expressed as an m-vector $\boldsymbol{g}$,

$$\boldsymbol{g} = (g_1, g_2, \ldots, g_m)^\top, \tag{5.5}$$

where $g_j = \Pr\{\theta = \theta^{(j)}\}$. Define

$$f_{ij} = \Pr\left\{x = x^{(i)} \mid \theta = \theta^{(j)}\right\}, \tag{5.6}$$

and let $\boldsymbol{M}$ be the $n \times m$ matrix $\{f_{ij}\}$. The marginal density $f(x)$ is an n-vector

$$\boldsymbol{f} = (f_1, f_2, \ldots, f_n)^\top, \qquad \text{where } f_k = \Pr\left\{x = x^{(k)}\right\},$$

with $\boldsymbol{f}$ given by the matrix product

$$\boldsymbol{f} = \boldsymbol{M}\boldsymbol{g}. \tag{5.7}$$

The observed data $x_1, x_2, \ldots, x_N$ is a random sample from $\boldsymbol{f}$. Sample space $\mathcal{X}$ being discrete, the counts

$$y_k = \#\left\{x_i = x^{(k)}\right\}, \qquad k = 1, 2, \ldots, n, \tag{5.8}$$

are sufficient statistics, with $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top$ having multinomial distribution

$$\boldsymbol{y} \sim \text{Mult}_n(N, \boldsymbol{f}) \tag{5.9}$$

(N draws on n categories with probability vector $\boldsymbol{f}$).

Parametric g-modeling assumes that $\boldsymbol{g}$ is a smooth function of a p-dimensional parametric vector α,

$$\{\boldsymbol{g}(\alpha), \alpha \in \mathcal{A}\}, \tag{5.10}$$

$\mathcal{A}$ a subset of $\mathcal{R}^p$; α produces the count vector $\boldsymbol{y}$ according to

$$\alpha \longrightarrow \boldsymbol{g}(\alpha) \longrightarrow \boldsymbol{f}(\alpha) = \boldsymbol{M}\boldsymbol{g}(\alpha) \longrightarrow \boldsymbol{y} \sim \text{Mult}_n\left(N, \boldsymbol{f}(\alpha)\right). \tag{5.11}$$

Maximizing the multinomial log likelihood

$$\text{loglik}(\alpha) = \sum_{k=1}^{n} y_k \log\left(f_k(\alpha)\right) \tag{5.12}$$

gives MLE $\hat{\alpha}$ and $\hat{\boldsymbol{g}} = \boldsymbol{g}(\hat{\alpha})$; an estimate of the prior density.

Efron (2016) proposed taking $\boldsymbol{g}(\alpha)$ to be a p-parameter exponential family,

$$g_j(\alpha) = Q_j^\top \alpha - \phi(\alpha) \qquad (j = 1, 2, \ldots, m). \tag{5.13}$$

Here the vectors $Q_j^\top$ are the rows of an $m \times p$ structure matrix $\boldsymbol{Q}$ chosen by the g-modeler. An example is given next. Function $\phi(\alpha)$ normalizes $\boldsymbol{g}(\alpha)$ to sum to 1,

$$\phi(\alpha) = \log \sum_{j=1}^{n} \exp\left(Q_j^\top \alpha\right). \tag{5.14}$$

Model (5.10)–(5.13) was applied to the prostate cancer data of Section 4, now with $z_1, z_2, \ldots, z_N$, $N = 6033$, playing the role of the x_i (5.1) and $z_i \sim \mathcal{N}(\theta_i, 1)$ as in (4.16). $\mathcal{X}$ equaled $\{-4.4, -4.2, \ldots, 5.2\}$ ($n = 49$), the bin centers in Figure 5, with the counts y_k the heights of the histogram bars. $\mathcal{T}$ was $\{-4.2, -4.0, \ldots, 4.2\}$, $m = 43$. The structure matrix $\boldsymbol{Q}$ was chosen to be

$$\underset{49\times 43}{\boldsymbol{Q}} = \left(I_0, \text{ns}(\mathcal{T}, 5)\right), \tag{5.15}$$

$I_0 = (0, 0, \ldots, 1, 0, 0, \ldots, 0)^\top$, the 1 at $\theta^{(22)} = 0$, and $\text{ns}(\mathcal{T}, 5)$ the R language natural spline function having five degrees of freedom (from the R package `splines`).

With this choice of $\boldsymbol{Q}$, (5.13) represents a "spike and slab" prior, with the spike at $\theta = 0$ accommodating the large number of null genes (zero effect size), while $\text{ns}(\mathcal{T}, 5)$ allows for smooth modeling of the non-null effects. Figure 8 shows the results: $\hat{\boldsymbol{g}}$ put probability 0.83 on $\theta = 0$, with most of

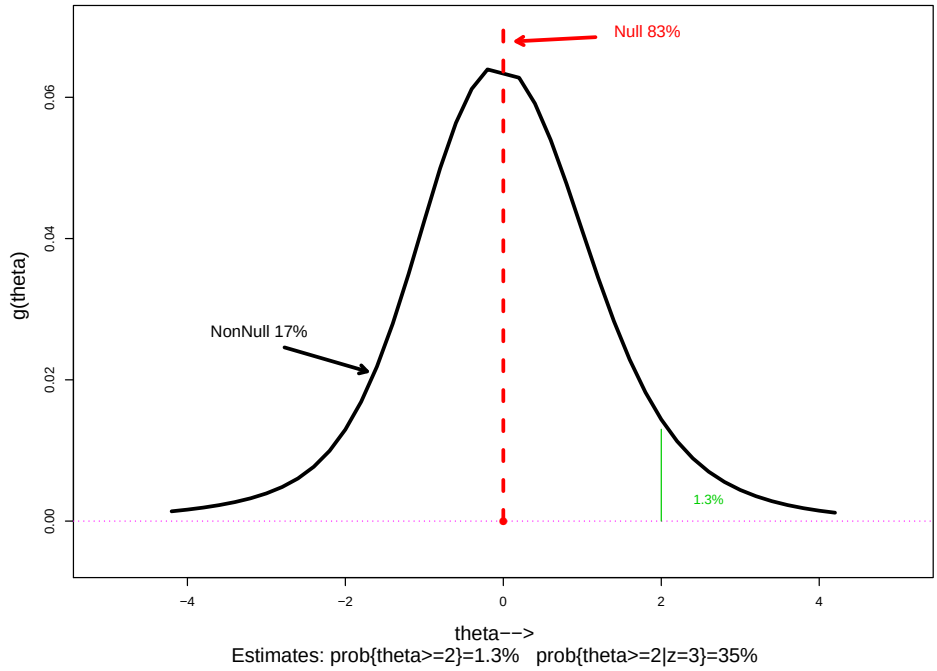

Figure 8: Prostate study: estimated prior $g(\theta)$; g-modeling $\boldsymbol{Q} = \text{ns}(\theta, 5)+$ atom at 0; estimated null probability 83%.

the 0.17 slab not very far from 0. For example, $\widehat{\Pr}\{\theta \geq 2\}$ was only 0.013. Taking $\hat{\boldsymbol{g}}$ literally allows for Bayesian estimates, e.g.,

$$\widehat{\Pr}\{\theta \geq 2 \mid z = 3\} = 0.35. \tag{5.16}$$

Section 2 of Efron (2016) provides the details for maximum likelihood estimation in model (5.13). Some regularization is necessary for stability; the regularization constant c_0 of Efron (2016, Sect. 3) was 0.5 in Figure 8.

g-modeling is inherently less straightforward than f-modeling. A low-dimensional parametric model, like that used for Figure 8, amounts to a bet on the form of $g(\theta)$. In (5.15) the bet is that the non-null prior distribution can be adequately approximated by a fifth-degree natural spline. The contribution by Koenker and Giu to the discussion of Efron (2019) describes nonparametric g-modeling in its modern computational form, which requires less betting but doesn't easily permit estimates like (5.16).

f-modeling has dominated empirical Bayes applications. Ingenious techniques like those of Robbins, Good and Toulmin, and Tweedie have allowed

certain special problems to be solved directly in terms of $f(x)$, rendering g-modeling unnecessary. Venturing beyond the class of "special problems" may lie in empirical Bayes' future, in which case g-modeling will have its day.

Whether or not that day arrives, g-models are the only game in town for analyses like those in Figure 8 or (5.16). Even on f-modeling's home ground, g-model estimates can be more efficient; see Tables 1 and 2 in Section 5 of Efron (2019). The computations in Figure 8 used the R package `deconvolveR`, fully described in Narasimhan and Efron (2020). g-modeling can be thought of as the computer-intensive child of Fisher's gamma model, just as the f-modeling methods of Section 4 are descendents of the James–Stein rule.

6 Relevance

Previously we have discussed empirical Bayes methods in terms of two dichotomies:

1. parametric or non-parametric
2. f-modeling or g-modeling

A third dichotomy concerns the use to which the methods are aimed,

3. ensemble or individual

(where "ensemble" would be the equivalent of "population" in a survey sampling framework). Ensemble methods hope to do well according to some overall measure of accuracy, as for instance EASE $= E\{\sum(\hat{\theta}_i - \theta_i)^2/N\}$ (3.18). "Individual" refers to inferences for a specific case, for example interpreting (5.16) to say that $\widehat{\Pr}\{\theta_i > 2 \mid z_i = 3\}$ applies to a specific gene having $z_i = 3$ in the prostate cancer study.

A fourth dichotomy is:

4. frequentist or Bayesian

Can empirical Bayes procedures be frequentist? The answer is yes, and in fact most of the literature has been approached from a frequentist point of view. The minimization of EASE in Section 3 is a frequentist criterion, with the prior $g(\theta)$ playing no role in the analysis except to suggest the form of the rule $\bar{g}(\theta)$ (3.20). Robbins' used the term *compound Bayes* to distinguish the Oracle Bayes situation, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_N)$ is fixed,

from *empirical Bayes*, where we truly believe $\theta_i \sim g(\cdot)$—but without much difference in his methodology.

Ensemble loss functions like $\sum(\hat{\theta}_i - \theta_i)^2$ are conducive to frequentism. Bayesian considerations become crucial when empirical Bayes methods are applied to individual inferences. Suppose the two towers in Figure 3 concern an insurance company where the right tower represents automobile claims and the left tower flood damages. The 40% savings of the empirical Bayes rule $\hat{e}(x)$ would still be relevant to the company's bottom line, but would be inappropriate if applied to any one individual's automobile or flood premium.

Similar considerations apply to the 18 baseball players in Table 6, Section 3. Baseball fans know that Clemente was the best hitter of his generation. Putting him in a James–Stein analysis with 17 less accomplished hitters almost guarantees an underestimate of his true ability. James–Stein shrinkage, Figure 3, is a winning tactic for minimizing total squared error loss, but risks "sacrificing the extremes for the sake of the center."

The question here is one of *relevance*. How relevant are the other 17 players to any one player's estimation? Frequentist estimates, like the MLE column in Table 6, base inference for each case on just its own data. Empirical Bayes methods share data across individuals, for the benefit of overall performance but at possible peril to unusual individuals. Bayes estimators have the same property — though usually concealed by a tacit assumption that all θ_i's obtained from the prior $g(\theta)$ are relevant to each other — but relevance is easier to worry about in an empirical Bayes setting, where we can see the "prior" being constructed from individual cases.

Figure 9 illustrates an example of individual inferences in an empirical Bayes setting: $n = 51$ values $x_0, x_1, x_2, \ldots, x_5$ have been observed from the empirical Bayes model

$$\theta_i \sim g(\theta) \quad \text{and} \quad x_i \mid \theta_i \sim \mathcal{N}(\theta_i, 1), \tag{6.1}$$

$i = 0, 1, 2, \ldots, 50$, with $g(\cdot)$ unknown and the θ_i unobserved. We wish to assess the posterior distribution of θ_0 given $x_0 = 5$ and $\boldsymbol{x} = (x_1, x_2, \ldots, x_{50})$. We are specifically interested in θ_0, and not some omnibus criterion such as EASE.

g-modeling[9] was used to produce an estimate $\hat{g}(\theta)$ for $g(\theta)$, the green dotted curve in Figure 10. The likelihood function for θ_0, $\varphi(\theta_0 - 5)$, is

[9] Using a natural spline with five degrees of freedom; a correction for the small sample size, $N = 50$, was made; see Section 6 of Efron (2019).

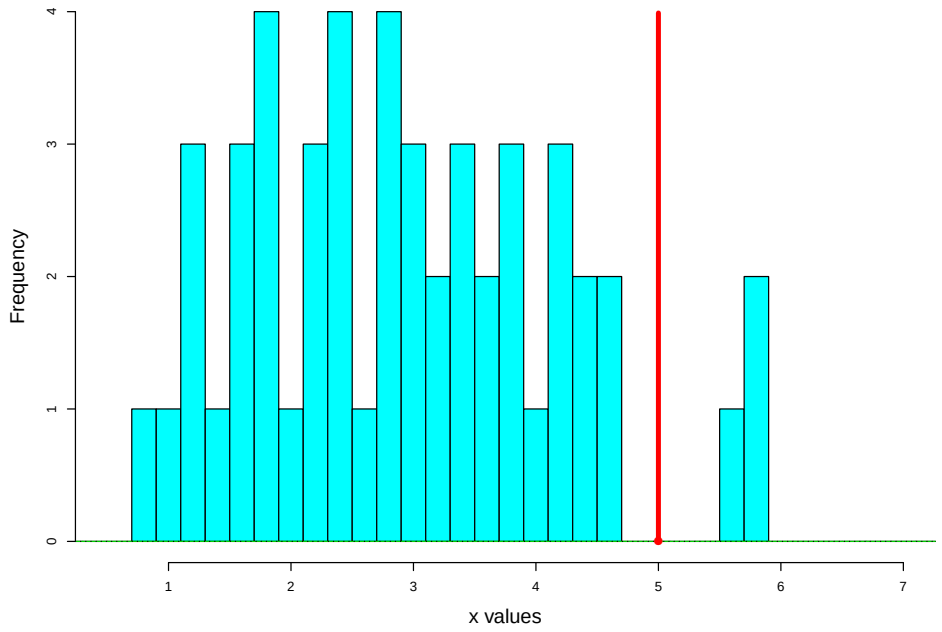

Figure 9: $x_0 = 5$ and 50 other values x_i are observed from model (6.1), shown as blue shaded histograms. What can we say about θ_0?

shown by the red dashed curve. Bayes rule then gives the posterior density

$$\hat{g}(\theta_0 \mid x_0, \boldsymbol{x}) = c\varphi(\theta_0 - 5)\hat{g}(\theta_0), \tag{6.2}$$

shown in solid black. The Bayesian effects are quite strong here: $\hat{g}(\theta_0 \mid x_0, \boldsymbol{x})$ reaches its maximum at $\theta_0 = 3.8$, compared to the standalone frequentist estimate $\hat{\theta}_0^{\mathrm{MLE}} = 5$; (6.2) puts only 18% of its mass above 5, compared to 50% for $\varphi(\theta_0 - 5)$. All of this requires us to take the Bayesian assumption $\theta_i \sim g(\theta)$ literally and not just as the hint it was in the EASE calculations of Section 3.

With only $N = 50$ values x_i to work with, one might worry about the accuracy of $\hat{g}(\theta_0 \mid x_0, \boldsymbol{x})$. The true prior $g(\theta)$ used to generate the x_i was $\mathrm{Gamma}_9/3$ (gamma with 9 degrees of freedom, divided by 3) and at least in this case, $\hat{g}(\theta_0 \mid x_0, \boldsymbol{x})$ is a reasonable approximation to the true posterior density $g(\theta_0 \mid x_0, \boldsymbol{x})$.

In a genuine (unsimulated) setting, a more pressing worry might concern the relevance of $x_1, x_2, \ldots, x_{50}$ to the estimation of θ_0. Empirical Bayes

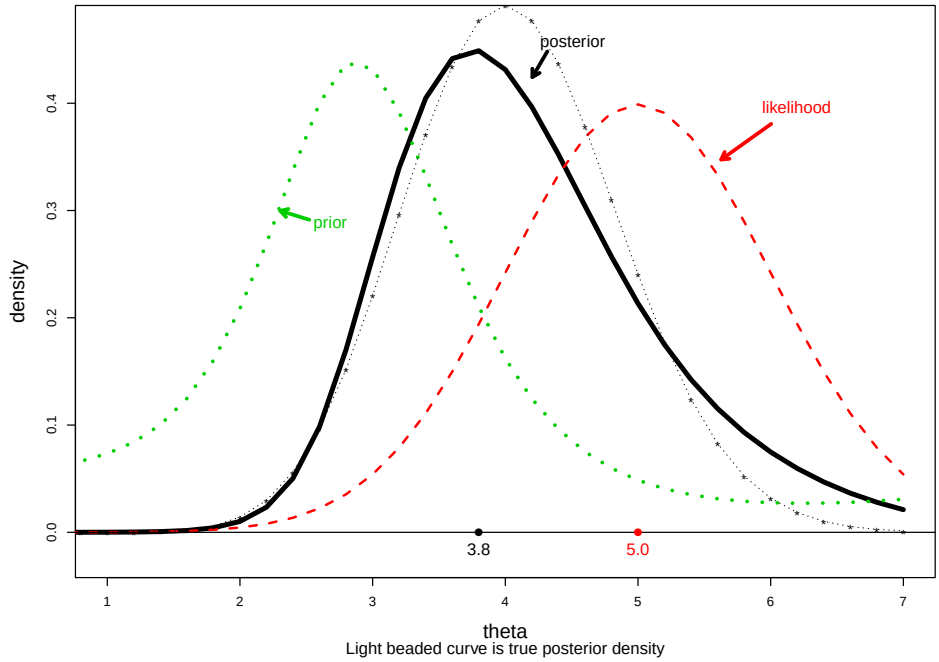

Figure 10: Empirical Bayes posterior inference for θ_0: likelihood (dashed red), g-model prior (dotted green), posterior density (solid black).

considerations have moved the point estimate from 5.0 down to 3.8, more than a full standard error from the MLE. Do we trust the empirical Bayes connection enough for such a drastic change? In a James–Stein context, Efron and Morris (1972) proposed capping the distance of the empirical Bayes estimate from the MLE, sacrificing some of the savings of $\hat{\theta}^{\mathrm{JS}}$ in order to protect outliers like Clemente from overshrinkage.

A standard inference problem might begin with data

$$x_{01}, x_{02}, \ldots, x_{0n} \overset{\text{iid}}{\sim} f(x \mid \theta_0), \tag{6.3}$$

in which case there is no question of the relevance of each observation to the estimation of θ_0. In the empirical Bayes model (6.1), the "other" observations x_i are less directly linked to θ_0; they have the same "grandparent" as $x_0 \sim f(x \mid \theta_0)$ through the common prior $g(\theta)$ but not the same parent θ_0, which is to say they are cousins to x_0 rather than siblings.

Specific relevance applications appear in Efron and Morris (1972), Efron (2010, Sect. 10.3), and Efron (2019, Sect. 7). A different methodology is

explored in Mukhopadhyay and Wang (2020). As a topic area, relevance is underdeveloped in the empirical Bayes literature, even though it can be an obvious concern in situations like that of the baseball players where the individual cases have evocative identities.

References

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B 57*(1), 289–300.

Brown, L. D. and E. Greenshtein (2009). Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *Ann. Statist. 37*(4), 1685–1704.

Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, Volume 1 of *Institute of Mathematical Statistics Monographs*. Cambridge: Cambridge University Press.

Efron, B. (2016). Empirical Bayes deconvolution estimates. *Biometrika 103*(1), 1–20.

Efron, B. (2019). Bayes, oracle Bayes and empirical Bayes. *Statist. Sci. 34*(2), 177–235. with commentary and a rejoinder by the author.

Efron, B. and C. Morris (1972). Empirical Bayes on vector observations: An extension of Stein's method. *Biometrika 59*, 335–347.

Efron, B. and R. Thisted (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika 63*(3), 435–447.

Fisher, R., A. Corbet, and C. Williams (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol. 12*, 42–58.

Good, I. and G. Toulmin (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika 43*, 45–63.

James, W. and C. Stein (1961). Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. Prob., Vol. I*, pp. 361–379. Berkeley: University of California Press.

Jiang, W. and C.-H. Zhang (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist. 37*(4), 1647–1684.

Kiefer, J. and J. Wolfowitz (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist. 27*, 887–906.

Koenker, R. and I. Mizera (2014). Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *J. Amer. Statist. Assoc. 109*(506), 674–685.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixed distribution. *J. Amer. Statist. Assoc. 73*(364), 805–811.

Lindsey, J. (1974). Construction and comparison of statistical models. *J. Roy. Statist. Soc. Ser. B 36*, 418–425.

Mukhopadhyay, S. and K. Wang (2020). On the problem of relevance in statistical inference. *arXiv e-prints*, arXiv:2004.09588.

Narasimhan, B. and B. Efron (2020). deconvolveR: A g-modeling program for deconvolution and empirical Bayes estimation. *J. Statist. Software 94*(11), 1–20.

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*, Berkeley and Los Angeles, pp. 157–163. University of California Press.

Spevack, M. (1968). *A Complete and Systematic Concordance to the Works of Shakespeare, Vols 1–6*. Hildesheim: George Olms.

Strawderman, W. E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *Ann. Math. Statist. 42*(1), 385–388.

Thisted, R. and B. Efron (1987). Did Shakespeare write a newly-discovered poem? *Biometrika 74*(3), 445–455.