BAYES, ORACLE BAYES, AND EMPIRICAL BAYES

By

Bradley Efron

Technical Report No. 2017-12
November 2017

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

BAYES, ORACLE BAYES, AND EMPIRICAL BAYES

By

Bradley Efron
Stanford University

Technical Report No. 2017-12
November 2017

**This research was supported in part by
National Science Foundation grant DMS 1608182.**

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

http://statistics.stanford.edu

Submitted to Statistical Science

Bayes, Oracle Bayes, and Empirical Bayes

Bradley Efron*

Stanford University

Abstract. This article concerns the Bayes and frequentist aspects of empirical Bayes inference. Some of the ideas explored go back to Robbins in the 1950s, while others are current. Several examples are discussed, real and artificial, illustrating the two faces of empirical Bayes methodology: "oracle Bayes" shows empirical Bayes in its most frequentist mode, while "finite Bayes inference" is a fundamentally Bayesian application. In either case, modern theory and computation allow us to present a sharp finite-sample picture of what is at stake in an empirical Bayes analysis.

MSC 2010 subject classifications: Primary 62B10; secondary 62F20.
Key words and phrases: finite Bayes inference, g-modeling, relevance, empirical Bayes regret.

1. INTRODUCTION

Empirical Bayes is the newest addition to the statistician's arsenal of inferential methodologies. By now, though, new isn't very new. Robbins' 1951 introduction of compound decision procedures marks a starting point, with the name "empirical Bayes" attached in his 1956 paper. The resulting era has provided us with more than 65 years of experience and exploration. Zhang (2003) gives an excellent brief review of Robbins' work and subsequent developments.

Considering the enormous gains potentially available from empirical Bayes methods, the effects on statistical practice have been somewhat underwhelming. A paucity of appropriate data sets has been part of the bottleneck. To be effective, empirical Bayes techniques require large numbers of parallel estimation or testing problems. Modern scientific technology excels in this direction, but before the introduction of microarrays in the 1990s, large-scale parallel inference problems were thin on the ground. The big data era should be a favorable one for empirical Bayes applications.

That being said, more data by itself might not fully open the floodgates. Empirical Bayes has suffered from a philosophical identity problem. Not firmly attached to either frequentism or Bayesianism, expositions of empirical Bayes typically hover uncertainly around the middle. In practice, empirical Bayes analysis employs both frequentist and Bayesian inferential methods. The main purpose of this paper is to clarify its dual nature. The basic ideas go back to the 1950s, but

Sequoia Hall, 390 Serra Mall, Stanford, CA 94305-4020; (e-mail: brad@stat.stanford.edu)

*Supported in part by National Science Foundation award DMS 1608182

substantial improvements in theory — and enormous improvements in computation — enable a sharper picture to emerge. A second purpose is to review some of the current technology and show it in action, with an emphasis on accurate finite-sample performance.

We will work in the following simplified framework: unobserved parameters θ_i have each independently generated an observation x_i according to a known probability kernel $p(x \mid \theta)$,

$$x_i \overset{\text{ind}}{\sim} p(x_i \mid \theta_i), \qquad i = 1, 2, \ldots, N. \tag{1}$$

Normal and Poisson distributions will be featured, $x_i \sim \mathcal{N}(\theta_i, 1)$ and $x_i \sim \text{Poi}(\theta_i)$, these being the most familiar and also the most amenable choices. It is desired to estimate the θ's. Robbins' key idea, and the launching point for empirical Bayes theory, is that the entire data set $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$ can profitably be employed in the estimation of each θ_i.

Section 2 introduces "oracle Bayes", as in Jiang and Zhang (2009), an artificial construction we will use here to emphasize the frequentist side of empirical Bayes applications. Later examples, both genuine and simulated, develop the Bayesian side of the story, a salient difference being whether the statistician is interested in individual inferences as opposed to some omnibus measure of accuracy for the entire vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_N)$. "Finite Bayes", Section 6, makes the individual inference problem explicit.

Empirical Bayes procedures typically add the assumption that the parameters θ_i in (1) have been independently drawn from some hidden prior density

$$\theta_i \overset{\text{ind}}{\sim} g(\theta), \qquad i = 1, 2, \ldots, N. \tag{2}$$

This raises the fundamental question of estimating $g(\cdot)$ from the observed data set $\boldsymbol{x}$. Nonparametric estimates are available (Laird, 1978) but here we will emphasize parametric modeling as in Efron (2016). (Section 6 includes some comments on nonparametric methods.)

The simplest case, where $g(\theta)$ is assumed to be normal, relates to the James–Stein estimator. Morris (1983) provided a normality-based theory of empirical Bayes confidence intervals. A more general but less exact approach to posterior intervals is discussed in Section 6, where the *Type 3 bootstrap* methodology of Laird and Louis (1987) plays a role. Posterior interval inference emphasizes the Bayesian side of empirical Bayes theory.

The marginal density $f(x)$ obtained from (1)–(2),

$$f(x) = \int_{\mathcal{T}} g(\theta) p(x \mid \theta)\, d\theta, \tag{3}$$

$\mathcal{T}$ the space of possible θ values, is central to empirical Bayes procedures, since $(x_1, x_2, \ldots, x_N)$ is no more than a random sample from $f(\cdot)$. In certain cases, and in fact in most of the familiar empirical Bayes applications, *only* $f(\cdot)$ need be estimated, thus avoiding the difficult deconvolution problems of estimating $g(\cdot)$. This is true for the oracle Bayes setup of Section 2. Both *f-modeling* and *g-modeling* — in the terminology of Efron (2014), that is, modeling $f(x)$ or $g(\theta)$ — are discussed in what follows, the latter inherently more attuned to the Bayesian side of empirical Bayes.

Most of the methodology reported in this paper is not new. Technical matters will mostly be deferred to the remarks of Section 8, clearing the way for a broad discussion of the Bayesian and frequentist aspects of empirical Bayes applications.

2. ORACLE BAYES

Suppose we observe a normal version of model (1),

$$x_i \overset{\text{ind}}{\sim} \mathcal{N}(\theta_i, 1), \qquad i = 1, 2, \ldots, N, \tag{4}$$

and use the data set $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$ to form estimates $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_N$, our goal being to minimize the expected average mean square error (ASE)

$$\text{ASE} = E_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{N} \left(\hat{\theta}_i - \theta_i\right)^2 \Big/ N \right\}. \tag{5}$$

The expectation here is over model (4), with the θ_i's fixed.

Using the maximum likelihood estimates (MLEs) $\hat{\theta}_i = x_i$ yields

$$\text{ASE}_{\text{MLE}} = 1. \tag{6}$$

However, a friendly Oracle has told us the *order statistic* of the true θ_i values, that is, their ordered values from smallest to largest

$$\boldsymbol{\theta}_{\text{ord}} = \left\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(N)}\right\}, \tag{7}$$

but not which observation x goes with which θ, allowing us to do better. The oracle Bayes setup (4)–(5) is pursued in Jiang and Zhang (2009), where sharp asymptotic properties of empirical Bayes procedures are developed.

Let $\bar{g}(\theta)$ denote the discrete density putting probability $1/N$ on each point $\theta^{(i)}$,

$$\bar{g}(\theta) = \sum_{i=1}^{N} \delta\left(\theta - \theta^{(i)}\right) \Big/ N, \tag{8}$$

$\delta(\cdot)$ the delta function at zero. Thanks to the Oracle we can compute $e_{\bar{g}}(x)$, the Bayes posterior expectation of θ given x, for prior $\bar{g}(\cdot)$,

$$e_{\bar{g}}(x) = \sum_{i=1}^{N} \theta^{(i)} \phi\left(x - \theta^{(i)}\right) \Big/ \sum_{i=1}^{N} \phi\left(x - \theta^{(i)}\right), \tag{9}$$

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

$$\hat{\theta}_i = e_{\bar{g}}(x_i) \tag{10}$$

will beat $\text{ASE}_{\text{MLE}} = 1$. A standard argument shows that the resulting ASE is the squared-error Bayes risk for estimating a single θ from $x \sim \mathcal{N}(\theta, 1)$, given prior $\bar{g}(\theta)$.

In the example of Figure 1, $\boldsymbol{\theta}_{\text{ord}}$ comprises $N = 1500$ values located in "two towers", 500 between -1.7 and -0.7, and 1000 between 0.7 and 2.7, as shown by the solid red histogram.. The black dashed histogram indicates 1500 x_i values

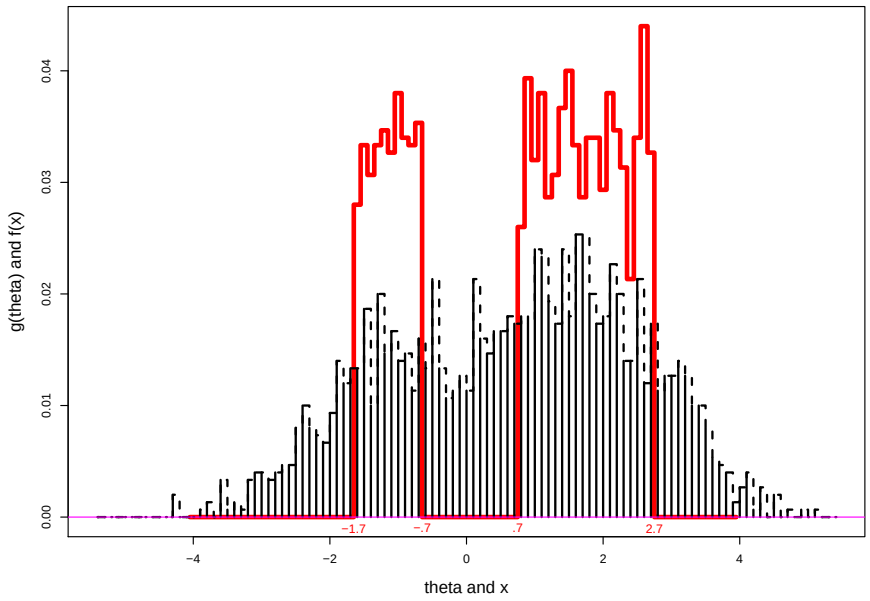

FIG 1. *Two towers example:* $N = 1500$ *parameters* θ_i *are known to follow the Oracle's solid red histogram, 500 in the left tower, 1000 in the right. We observe* $x_i \overset{ind}{\sim} \mathcal{N}(\theta_i, 1)$ *for* $i = 1, 2, \ldots, 1500$ *(dashed black histogram) and wish to estimate* $(\theta_1, \theta_2, \ldots, \theta_{1500})$*. Using the Oracle's information reduces ASE by more than 40%; empirical Bayes methods allow us to do almost as well without the Oracle's help.*

from a particular realization of (4). (The position of the towers was chosen to make the marginal density of the x_i's just barely bimodal.)

Formula (28) of Section 3, applied to the oracle Bayes estimation rule (9), gave expected average squared error

$$\text{ASE}_{\bar{g}} = 0.563 \tag{11}$$

for the two towers prior $\bar{g}$. Compared with $\text{ASE}_{\text{MLE}} = 1.0$, the Oracle's information has saved us more than 40% of the average estimation error.

Of course, real-life oracles are in short supply. This is where empirical Bayes makes its entrance: the full data set $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$ is used to form an estimate $\hat{g}(\cdot)$ of the empirical density $\bar{g}(\cdot)$, from which we calculate its Bayes posterior expectation,

$$e_{\hat{g}}(x) = \int_{\mathcal{T}} \theta \hat{g}(\theta)\phi(x-\theta)\, d\theta \Big/ \int_{\mathcal{T}} \hat{g}(\theta)\phi(x-\theta)\, d\theta, \tag{12}$$

yielding estimates $\hat{\theta}_i = e_{\hat{g}}(x_i)$.

These cannot be as accurate as the oracle Bayes estimates $e_{\bar{g}}(x_i)$, but the *empirical Bayes regret* may be surprisingly small. The g-modeling methods of Table 1 in Section 4 give

$$\text{EBregret} = \text{ASE}_{\hat{g}} - \text{ASE}_{\bar{g}} = 0.008, \tag{13}$$

so $\text{ASE}_{\hat{g}} = 0.571$ is still more than 40% less than ASE_{MLE}. Effectively, we have fashioned our own oracle from the data. The EBregret formula in Section 4 suggests regret declining as $1/N$ with sample size. Reducing N from 1500 to 150 increases EBregret to about 0.08, giving $\text{ASE}_{\bar{g}} \doteq 0.64$, still much less than ASE_{MLE}.

All of these inferences are frequentist in nature. First of all, ASE is a frequentist criterion. Moreover, the Bayesian assumption $\theta_i \overset{\text{ind}}{\sim} g(\theta)$ (2) plays only a motivational role behind $\bar{g}$ or $\hat{g}$, and is irrelevant to their application. In Figure 1, for example, the θ's of the left tower might relate to traffic accidents and those of the right to flood damage claims, or there might be dozens of other θ types among the 1500. Even so, the 40% reduction in ASE could still be meaningful, say to an insurance actuary planning next year's rates.

The Bayesian side of empirical Bayes emerges when we take an estimated prior $\hat{g}(\theta)$ seriously for the inference of an individual parameter θ_i, perhaps through the posterior density

$$\hat{g}(\theta_i \mid x_i) = \hat{g}(\theta_i)p(x_i \mid \theta_i)/\hat{f}(x_i), \tag{14}$$

$\hat{f}$ the marginal density (3) corresponding to $\hat{g}$. Now we wouldn't want to mix traffic accidents with flood claims.

This brings up the question of *relevance*: what cases can legitimately be combined in an empirical Bayes analysis? An example of the tension between omnibus accuracy — that 40% reduction — and individual relevance will be taken up in Section 7 in the context of an fMRI study. Efron and Morris (1972) considered relevance questions in terms of the James–Stein estimator, perhaps the best-known empirical Bayes construction. See also Chapter 7 of Efron and Hastie (2016). Section 6 here, on "finite Bayes", directly examines the estimation of a single θ_i of interest within an empirical Bayes framework.

3. BAYES RISK AND REGRET

The oracle Bayes model looks more familiar if we let the number of cases N go to infinity in (1)–(2). Then $\bar{g}(\theta)$ (8) converges to $g(\theta)$, and the inference for any one θ_i follows from the usual single-case Bayesian setup,

$$\theta \sim g(\theta) \quad \text{and} \quad x \mid \theta \sim p(x \mid \theta). \tag{15}$$

In other words, standard Bayes *is* oracle Bayes, where past experience has provided the oracle.

The next two paragraphs review Bayesian estimation of θ for model (15). We assume that x given θ is unbiased with variance $V(\theta)$,

$$x \mid \theta \sim (\theta, V(\theta))\,, \tag{16}$$

and denote the posterior expectation and variance of θ given x by

$$\theta \mid x \sim (e_g(x), v_g(x))\,; \tag{17}$$

$\hat{\theta} = e_g(x)$ is the Bayes estimate of θ under squared error loss. Its overall Bayes risk $\mathcal{R}_g$ is

$$\begin{aligned} \mathcal{R}_g &= E\left\{\left(\hat{\theta}-\theta\right)^2\right\} = \int_{\mathcal{T}}\int_{\mathcal{X}} \left(e_g(x)-\theta\right)^2 p(x \mid \theta)g(\theta)\; d\theta \\ &= \int_{\mathcal{X}} v_g(x)f(x), \end{aligned} \tag{18}$$

where $f(x)$ is the marginal density (3) and $\mathcal{X}$ is the sample space of the observations x.

Now suppose that instead of $e_g(x)$, we must use some other estimate $\hat{\theta} = \hat{e}(x)$. This increases the overall risk versus prior $g(\theta)$ to

$$\begin{aligned}\mathcal{R}(g,\hat{e}) &= E\left\{(\hat{e}(x)-\theta)^2\right\} = E\left\{(\hat{e}(x)-e_g(x)+e_g(x)-\theta)^2\right\}\\ &= \mathcal{R}_g + E\left\{(\hat{e}(x)-e_g(x))^2\right\},\end{aligned} \tag{19}$$

so our *regret* is

$$\mathcal{R}(g,\hat{e}) - \mathcal{R}_g = \int_{\mathcal{X}} (\hat{e}(x)-e_g(x))^2 f(x)\, dx. \tag{20}$$

The unbiased estimate $\hat{e}(x) = x$ has Bayes risk

$$\mathcal{R}(g,\hat{e}) = \int_{\mathcal{T}} V(\theta) g(\theta)\, d\theta \equiv V_g, \tag{21}$$

the average variance. Formula (20) provides a convenient expression for $\mathcal{R}_g$ that we will use later.

LEMMA 3.1.

$$\mathcal{R}_g = V_g - \int_{\mathcal{X}} (x-e_g(x))^2 f(x)\, dx. \tag{22}$$

The difference between x and $e_g(x)$ determines the amount of Bayesian savings available.

Tweedie's formulas (Efron, 2011) provide useful expressions for $e_g(x)$ and $v_g(x)$. Suppose $p(x \mid \theta)$ in (15) is a one-parameter exponential family,

$$p(x \mid \theta) = e^{\theta x - \psi(\theta)} p_0(x), \tag{23}$$

with natural parameter θ, sufficient statistic x, normalizing function $\psi(\theta)$, and base density $p_0(x)$. Let $l(x)$ be the log of the marginal density $f(x)$ (3) and $l_0(x) = \log p_0(x)$. Tweedie's formulas give convenient expressions for $e_g(x)$ and $v_g(x)$ (17),

$$\begin{aligned} e_g(x) &= E\{\theta \mid x\} = \dot{l}(x) + \dot{l}_0(x),\\ v_g(x) &= \mathrm{Var}\{\theta \mid x\} = \ddot{l}(x) + \ddot{l}_0(x),\end{aligned} \tag{24}$$

the dots indicating first and second derivatives with respect to x. See Remark A of Section 8.

The normal case (4) has densities $p(x \mid \theta)$ equaling

$$e^{-(x-\theta)^2/2} \Big/ \sqrt{2\pi} = e^{\theta x - \theta^2/2} \phi(x), \tag{25}$$

so $p_0(x)$ in (23) is $\phi(x)$ and $l_0(x) = -x^2/2 - \log\sqrt{2\pi}$. Tweedie's formulas become

$$e_g(x) = x + \dot{l}(x) \quad \text{and} \quad v_g(x) = 1 + \ddot{l}(x). \tag{26}$$

(See Remark B of Section 8 for $x_i \sim \mathcal{N}(\theta_i, \sigma^2)$, σ^2 known.) From (18) we obtain the overall Bayes risk $\mathcal{R}_g$,

$$\mathcal{R}_g = \int_{-\infty}^{\infty} \left(1 + \ddot{l}(x)\right) f(x)\, dx = 1 - \int_{-\infty}^{\infty} \dot{l}(x)^2 f(x)\, dx, \tag{27}$$

the final expression obtained by integrating $\ddot{l}(x) = \ddot{f}(x)/f(x) - (\dot{f}(x)/f(x))^2$. It can also be written as

$$\mathcal{R}_g = 1 - \int_{-\infty}^{\infty} (x - e_g(x))^2\ dx \tag{28}$$

using (26), this being the same as Lemma 3.1 (22) since $V_g = 1$ in situation (4).

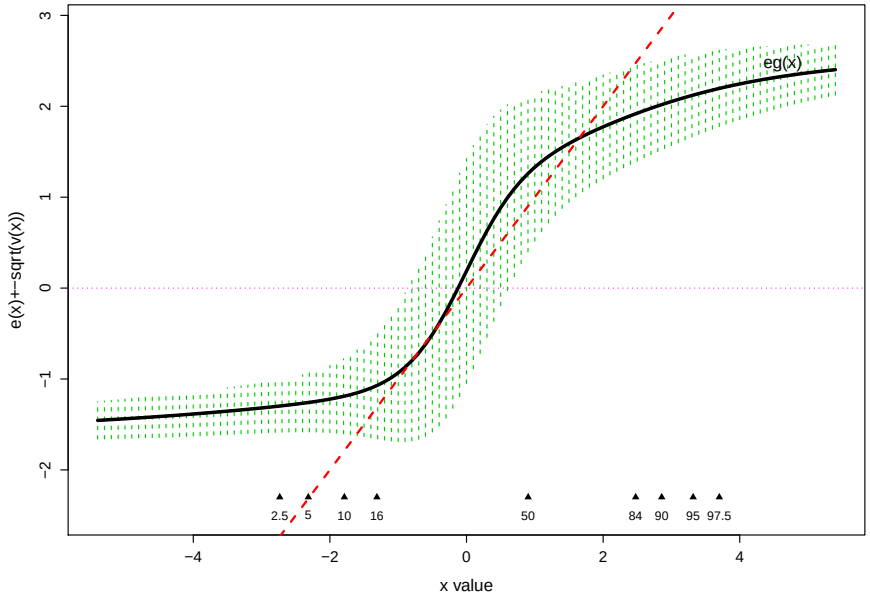

FIG 2. *Posterior expectation $e_{\bar{g}}(x)$ (9) for the oracle prior $\bar{g}(\cdot)$ of Figure 1; vertical dashed bars indicate ± one posterior standard deviation $v_{\bar{g}}(x)^{1/2}$. Dashed red line is main diagonal. Small triangles show the indicated percentiles for the marginal density $f_{\bar{g}}(x)$.*

Figure 2 shows $e_{\bar{g}}(x)$ (9), the posterior expectation $E\{\theta \mid x\}$ for the oracle prior (8). Numerical integration of formula (28) gives (11),

$$\mathrm{ASE}_{\bar{g}} = \mathcal{R}_{\bar{g}} = 0.563. \tag{29}$$

The dashed vertical green bars indicate Bayes posterior variability

$$e_{\bar{g}}(x) \pm v_{\bar{g}}(x)^{1/2}. \tag{30}$$

To restate a previous point, Figure 2 is a purely frequentist construction: it depends only on $\boldsymbol{\theta}_{\mathrm{ord}}$ (7) and not on any Bayesian assumptions regarding the θ_i's, such as (2). Assumption (2) becomes crucial if we use the figure for statements of posterior inference such as

$$\Pr\left\{\theta_i \in e_{\bar{g}}(x_i) \pm 1.96 v_{\bar{g}}(x_i)^{1/2} \mid x_i\right\} \doteq 0.95, \tag{31}$$

as discussed in Section 6.

4. *F*-MODELING AND LINDSEY'S METHOD

We would like to estimate the Bayes risk $\mathcal{R}_{\bar{g}}$, or ASE (5), from the observed data $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$, without the help of an oracle. Looking at Figure 1, a simple procedure suggests itself:

1. Estimate the marginal density $f(x)$ (3) by a smooth curve $\hat{f}(x)$ drawn through the bar tops of the black dashed histogram.
2. Estimate the conditional expectation $e_{\bar{g}}(x) = E\{\theta \mid x\}$ according to (26),

$$\hat{e}(x) = x + \frac{d}{dx}\log \hat{f}(x). \tag{32}$$

3. Estimate $\mathcal{R}_{\bar{g}}$ using Lemma 3.1 (22).

Step 1 is a definitional statement of *f-modeling*. Nonparametric or semiparametric techniques are available, but efficiency is crucial here. A parametric approach using *Lindsey's method*, as in Section 5.2 of Efron (2010), is particularly easy to implement. The sample space $\mathcal{X}$ is partitioned into K bins; for bin_k we compute the count y_k of observations it contains,

$$y_k = \#\{x_i \text{ in } \text{bin}_k\}, \tag{33}$$

and also its centerpoint $x_{(k)}$. Figure 1 hs $K = 109$ bins, each of width 0.10, with y_k proportional to the height of the black bars.

In the computations that follow, $\boldsymbol{f} = (f_1, f_2, \ldots, f_K)$ will represent a discrete probability distribution for observation x,

$$f_k = \Pr\{x \in \text{ bin}_k\}, \tag{34}$$

with $\bar{\boldsymbol{f}} = (\bar{f}_1, \bar{f}_2, \ldots, \bar{f}_K)$ denoting the marginal density induced by $\boldsymbol{\theta}_{\text{ord}}$ (7) and $\hat{\boldsymbol{f}} = (\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_K)$ the density corresponding to $\hat{f}(x)$; similarly, we write $\hat{\boldsymbol{e}} = (\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_K)$ for the vector of estimates (32) evaluated at the bin centers $x_{(k)}$.

Lindsey's method uses Poisson regression to estimate $f(x)$. The counts y_k are taken to be independent Poisson variates with expectations proportional to f_k,

$$y_k \overset{\text{ind}}{\sim} \text{Poi}(N \cdot f_k) \qquad \text{for } k = 1, 2, \ldots, K; \tag{35}$$

$\log \boldsymbol{f}$ is assumed to have a linear form

$$\log \boldsymbol{f} = \boldsymbol{M}\beta, \tag{36}$$

$\boldsymbol{M}$ a given $N \times p$ structure matrix and β an unknown p-dimensional parametric vector; finally $\hat{\boldsymbol{f}}$ is estimated by Poisson regression,

$$\hat{\boldsymbol{f}} = \texttt{glm(y} \sim \mathbf{M}\texttt{, poisson)\$ fit/N} \tag{37}$$

in R notation.

The three-step algorithm was carried out using the data from the black dashed histogram of Figure 1, with

$$\boldsymbol{M} = \text{ns}(\boldsymbol{x}_{()}, \text{df} = 7), \tag{38}$$

$\boldsymbol{x}_{()} = (x_{(1)}, x_{(2)}, \ldots, x_{(K)})$ the vector of bin centers and "ns" indicating natural splines, here invoked with 7 degrees of freedom. It gave estimated Bayes risk (22)

$$\hat{\mathcal{R}} = 1 - \sum_{k=1}^{K} \hat{f}_k (x_{(k)} - \hat{e}_k)^2 = 0.541. \tag{39}$$

Its actual ASE versus $\boldsymbol{\theta}_{\text{ord}}$ from the Oracle was, using (11) and (20),

$$\mathcal{R}(\bar{g}, \hat{e}) = \mathcal{R}_{\bar{g}} + \sum_{k=1}^{K} \bar{f}_k (\hat{e}_k - \bar{e}_k)^2 = 0.580. \tag{40}$$

So EBregret $= 0.580 - 0.563 = 0.017$.

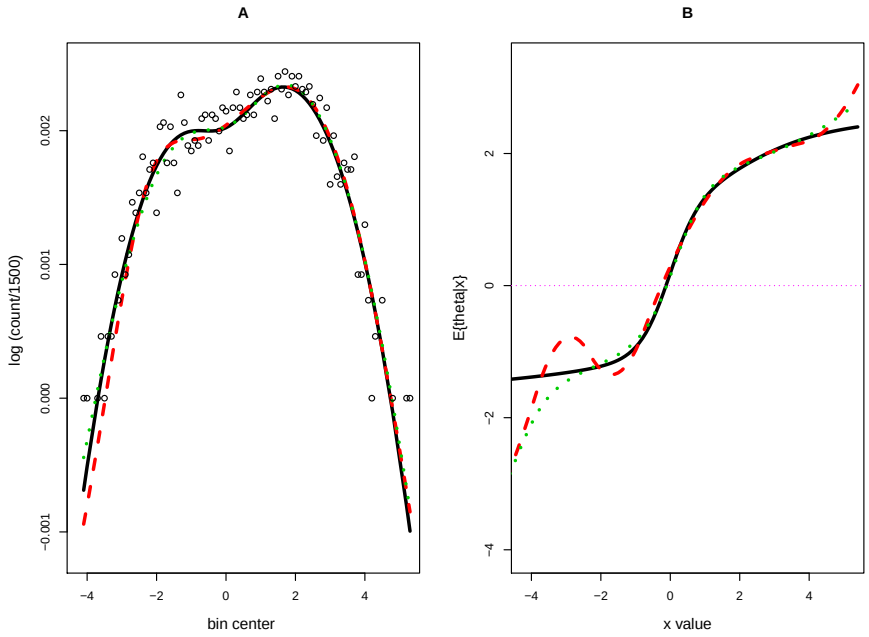

FIG 3. *Estimating Bayes risk (ASE) from the sample of $N = 1500$ observations in Figure 1.* ***Left panel:*** *Open circles $\log\{y_k/N\}$; black solid curve $\log\{\bar{f}_k\}$, true oracle marginal density; red dashed curve $\log\{\hat{f}_k\}$ from f-modeling.* ***Right panel:*** *Corresponding estimates of $e(x) = E\{\theta \mid x\}$; green dotted curves are from g-modeling.*

The fitting procedure is illustrated in the left panel of Figure 3: open circles plot the log counts versus bin centers (ignoring zeros),

$$\left(x_{(k)}, \log\{y_k/N\}\right), \qquad k = 1, 2, \ldots, K; \tag{41}$$

the black curve plots $\bar{f}_k$, the Oracle's true marginal density (11); and the red dashed curve plots $\hat{f}_k$, the estimated density from Lindsey's method (37).

It looks like a close fit, but going from $\hat{f}(x)$ to $\hat{e}(x)$ (32) (using finite differences of $\log \hat{f}_k$) exacerbates small errors, especially near the extreme values of x. This is seen in the right panel of Figure 3, where the true $\bar{e}(x)$ is compared with $\hat{e}(x)$. The error in (40), $\sum \bar{f}_k(\hat{e}_k - \bar{e}_k)^2$, is mitigated by the small values of $\bar{f}_k$ near the extremes, but is still substantial.

A second pair of estimates $\hat{f}(x)$ and $\hat{e}(x)$ are shown as the green dotted curves in Figure 3. These are based on *g-modeling* as described in Section 5, where

exponential family models are applied to $g(\theta)$ rather than $f(x)$. The prior $g(\theta)$ is hidden in empirical Bayes applications, which makes g-modeling inherently more involved than f-modeling, but often less noisy.

TABLE 1

Simulation study of 100 samples (4), $N = 1500$, *from* $\boldsymbol{\theta}_{\text{ord}}$ *in Figure 1. Estimated Bayes risk* $\hat{\mathcal{R}}$ (39) *and actual Bayes risk (ASE)* $\mathcal{R}(\bar{g}, \hat{e})$ (40) *computed using* f- *and* g-*modeling (both methods employed natural spline models with 7 degrees of freedom); g-modeling reduced EBregret by more than half. "Formula EBregret" used f-modeling as in Lemma 4.1* (47) *and g-modeling as in Remark F of Section 8. The value* EBregret $= 0.008$ *in* (13) *is from the entry 0.0082 here.*

	Estimated Bayes Risk		True Bayes Risk		True EBregret		Formula EBregret	
	f	g	f	g	f	g	f	g
mean	.557	.589	.581	.571	.0184	.0082	.0119	.0064
stdev	.024	.018	.007	.003	.0074	.0031	.0010	.0011

That is the case here. Table 1 reports on a simulation study in which 100 samples $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$, $N = 1500$, were drawn according to (4) with the θ values equaling $\boldsymbol{\theta}_{\text{ord}}$ in Figure 1 and the fitting done as in (38); $\hat{\mathcal{R}}$ and $\mathcal{R}(\bar{g}, \hat{e})$, (39) and (40), were computed for each sample, for both f- and g-modeling. The table lists means and standard deviations for the 100 trials; g-modeling was consistently less noisy and more accurate. In particular, the EBregret $\mathcal{R}(\bar{g}, \hat{e}) - \mathcal{R}_{\bar{g}}$ was halved by g-modeling.

In addition to estimating the Bayes risk (39) from the observed data $\boldsymbol{x}$, we might wish to estimate the empirical Bayes regret $\mathcal{R}(\bar{g}, \hat{e}) - \mathcal{R}_{\bar{g}}$,

$$\text{EBregret} = \sum_{k=1}^{K} \bar{f}_k (\hat{e}_k - \bar{e}_k)^2. \tag{42}$$

This is more difficult since regret is the *difference* of two risks. A useful but not fully dependable *delta method* formula is discussed next.

Let $\mu_k = N f_k$ so that $y_k \overset{\text{ind}}{\sim} \text{Poi}(\mu_k)$ for $k = 1, 2, \ldots, K$ in (35), or more succinctly,

$$\boldsymbol{y} \sim \text{Poi}(\boldsymbol{\mu}). \tag{43}$$

Poisson generalized linear models (GLMs) assume that the vector $\log(\boldsymbol{\mu}) = (\cdots \log(\mu_k) \cdots)$ is of the form

$$\log(\boldsymbol{\mu}) = \boldsymbol{M}\beta, \tag{44}$$

where $\boldsymbol{M}$ is a known $N \times p$ structure matrix and β is an unknown p-dimensional parameter vector. ($p = 8$ in (37)–(38), including the intercept term.)

Also let $\dot{\boldsymbol{M}}$ be the $N \times p$ matrix

$$\dot{\boldsymbol{M}} = \boldsymbol{D}\boldsymbol{M}, \tag{45}$$

where $\boldsymbol{D}$ is an operator that differentiates the rows of $\boldsymbol{M}$. For ease of application, if $\boldsymbol{x}_{()}$ is a regular grid of points with spacings Δ then we can take $\boldsymbol{D}$ to be the $N \times N$ matrix having kth row

$$(0, 0, \ldots, 0, -1/\Delta, 0, 1/\Delta, 0, \ldots, 0), \tag{46}$$

the nonzeros in places $k - 1$ and $k + 1$ (with modifications at $k = 1$ and K).

LEMMA 4.1. *A delta method estimate for EBregret is*

$$\widehat{\text{EBregret}} = \frac{1}{N}\,\text{trace}\left\{\left(\boldsymbol{M}'\,\text{diag}\left(\hat{\boldsymbol{f}}\right)\boldsymbol{M}\right)^{-1}\left(\dot{M}'\,\text{diag}\left(\hat{\boldsymbol{f}}\right)\dot{M}\right)\right\}, \tag{47}$$

$\text{diag}(\hat{\boldsymbol{f}})$ *the diagonal matrix with entries* $\hat{f}_k$.

A derivation is given in Remark B of Section 8. Lemma 4.1 approximates $(\hat{e}_k - \bar{e}_k)^2$ by an estimate of $\text{Var}(\hat{e}_k)$, ignoring bias. Bias, however, is a major factor in the example of Figure 1, where the smooth model (38) is poorly matched to the discontinuous two towers prior. For the 100 trials involved in Table 1, $\widehat{\text{EBregret}}$ from (47) averaged 0.0119, compared to 0.0184 for the true EBregret.

A less pathological situation is the *gamnormal* example featured in Section 6, where $\boldsymbol{\theta}_{\text{ord}}$ is determined by

$$\theta_i \overset{\text{ind}}{\sim} \text{Gamma}_9\,/3 \qquad \text{for } i = 1, 2, \ldots, N = 3200, \tag{48}$$

Gamma_9 a gamma variate with 9 degrees of freedom, and $x_i \overset{\text{ind}}{\sim} \mathcal{N}(\theta_k, 1)$ as before. It has oracle ASE $\mathcal{R}_{\bar{g}} = 0.489$.

TABLE 2

Simulation study of 100 samples (4)*,* $N = 3200$*, from fixed* $\boldsymbol{\theta}_{\text{ord}}$ *determined by* (48)*; true ASE* $\mathcal{R}_{\bar{g}} = 0.489$*; see Section 6. Both* f*- and* g*-modeling employed natural spline models with 5 degrees of freedom. Now* f*-modeling is more competitive, and the sample-based EBregret formulas are more accurate.*

	Estimated Bayes Risk		True Bayes Risk		True EBregret		Formula EBregret	
	f	g	f	g	f	g	f	g
mean	.481	.494	.496	.493	.0068	.0036	.0060	.0032
stdev	.014	.013	.005	.001	.0052	.0014	.0002	.0008

A simulation study similar to that in Table 1 was run for situation (48), with the results reported in Table 2. Here both f- and g-modeling relied on natural splines with 5 degrees of freedom. Now f-modeling was more competitive, though it still gave larger and more variable realizations of EBregret. Formula (47) averaged 0.0060 compared to 0.0068 for the average true regret. A data-based formula for estimating EBregret — which *does* include a bias term — is discussed in Section 8. It performed moderately well in Table 1 and Table 2.

5. POISSON OBSERVATIONS AND G-MODELING

The very earliest empirical Bayes papers — Fisher, Corbet and Williams (1943), Good and Toulmin (1956), Robbins (1956) — involved Poisson observations x_i,

$$\theta_i \overset{\text{ind}}{\sim} g(\theta), \quad x_i \overset{\text{ind}}{\sim} \text{Poi}(\theta_i) \qquad \text{for } i = 1, 2, \ldots, N. \tag{49}$$

Poisson data is more interesting than the normal case (4) in the sense that there is more than one obvious path to follow.

Robbins provided a notable Poisson formula for $e_g(x) = E\{\theta \mid x\}$,

$$e_g(x) = (x+1)f(x+1)/f(x), \tag{50}$$

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

$$f(x) = \int_{\mathcal{T}} p(x \mid \theta) g(\theta)\, d\theta, \tag{51}$$

$p(x \mid \theta) = e^{-\theta}\theta^x/x!$ for $x = 0, 1, 2, \ldots$. See for example, Chapter 6 of Efron and Hastie (2016). Similar reasoning gives the conditional variance $v_g(x) = \mathrm{Var}\{\theta \mid x\}$,

$$v_g(x) = e_g(x)\,(e_g(x+1) - e_g(x))\,. \tag{52}$$

Formulas (50) and (52) provide an impetus for f-modeling: in an empirical Bayes situation, where $g(\cdot)$ is unknown in (49), we need only estimate $f(\cdot)$ to approach the Bayes estimate and its risk.

TABLE 3

Corbet's butterfly data. After two years in Malaysia, Corbet had trapped 118 species just one time each, 74 species twice each, etc., $N = 501$ species in total. He asked Fisher to calculate how many new species would be seen if trapping continued for another year.

x	1	2	3	4	5	6	7	8	9	10	11	12
y	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	6	12	6	9	9	6	10	10	11	5	3	3

Corbet's butterfly data, Table 3, has a claim to being the initial vehicle for empirical Bayes analysis. Alexander Corbet, prominent naturalist, had been trapping butterflies in Malaysia (then Malaya) for two years in the early 1940s: 118 very rare species had been trapped just once each, 74 twice each, etc., as shown in the table,

$$y_x = \#\{\text{species having } x_i = x\} \tag{53}$$

for $x = 1, 2, \ldots, 24$. The total number in the table is $N = 501 = \sum y_x$. Corbet asked R.A. Fisher how many *new* species he could expect to see if he continued trapping for one more year. We will return to the answer at the end of this section.

We assume model (49), that species i is observed according to a Poisson distribution having expectation θ_i, but with an important modification: that x_i is only observed if it falls into

$$\mathcal{X} = \{1, 2, \ldots, 24\}; \tag{54}$$

that is, x_i follows a *truncated* Poisson distribution,

$$\theta_i \overset{\text{ind}}{\sim} g(\theta), \qquad x_i \overset{\text{ind}}{\sim} \mathrm{Poi}_{\mathcal{X}}(\theta_i), \tag{55}$$

$\mathrm{Poi}_{\mathcal{X}}(\theta)$ having density function

$$p(x \mid \theta) = e^{-\theta}\theta^x/(x!P_\theta) \qquad \text{for } x \in \mathcal{X}, \tag{56}$$

where $P_\theta = \sum_{\mathcal{X}} e^{-\theta}\theta^x/x!$. Truncation modifies the marginal density $f(x)$ and the effective prior density $g(\theta)$, but Robbins' formulas (50) and (52) remain valid as stated; see Remark D.

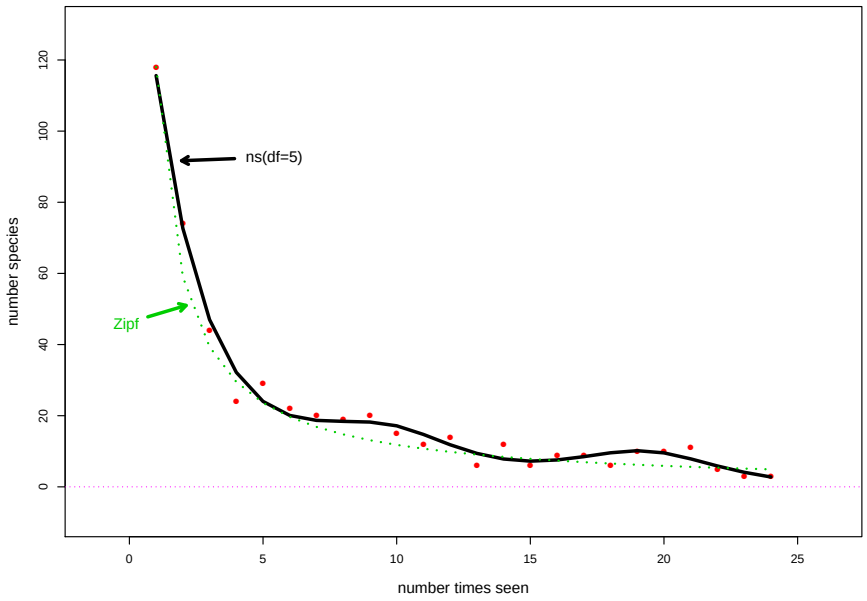

FIG 4. *Corbet's butterfly data. Red points are the* (x, y) *data from Table 3; solid black curve is natural spline Poisson regression fit, 5 degrees of freedom* (57). *Light green dashed curve follows Zipf's law:* $\hat{y}_x = y_1/x$.

The points in Figure 4 plot y versus x from Table 3. A smooth curve $N \cdot \hat{f}(x)$ has been fit to the points by Lindsey's method (37), using a natural spline model on $\mathcal{X}$ with five degrees of freedom,

$$\hat{\boldsymbol{f}} = \texttt{glm(y} \sim \texttt{ns(}\mathcal{X}\texttt{, df} = 5\texttt{), poisson)\$ fit/N}, \tag{57}$$

with k in (33) the same as x here; notice that the Poisson assumption in (57) is distinct from that in (49). The fit is excellent: chi-squared $= 12.2$ on $18 = 24 - 6$ degrees of freedom.

The famous (or notorious) *Zipf's law* predicts

$$y_x = y_1/x \qquad \text{for } x = 1, 2, \ldots, \tag{58}$$

plotted as the light dashed curve in Figure 4. This also fits reasonably well: chi-squared 28.1 on $23 = 24 - 1$ degrees of freedom, p-value 0.21. Zipf's law interacts in a surprising way with Robbins' formula (50): if $f(x)$ is proportional to $1/x$ then

$$e_g^{\text{Zipf}}(x) = x. \tag{59}$$

That is, *the Bayes estimate* $E\{\theta \mid x\}$ *is identical to the "MLE"* $\hat{\theta} = x$. (The quotes are a reminder that $\hat{\theta} = x$ is not exactly the MLE for a truncated Poisson distribution, a distinction ignored in the next paragraph.)

The Poisson family has variance $V(\theta) = \theta$ in (16), so that V_g (21) equals $\int_{\mathcal{T}} g(\theta)\theta \, d\theta$, the overall expectation of θ; this is the same as the marginal expectation of x, suggesting the estimate

$$\hat{V}_g = \bar{x} \tag{60}$$

for use in (22), equaling *6.60* for the butterfly data. Lemma 3.1 then gives Bayes risk

$$\mathcal{R}_g = 6.60 - \sum_{\mathcal{X}} f_x \left(x - e_g(x)\right)^2, \tag{61}$$

the second term, or *Bayes savings*, depending on the discrepancy between $e_g(x)$ and $x = e_g^{\text{Zipf}}(x)$.

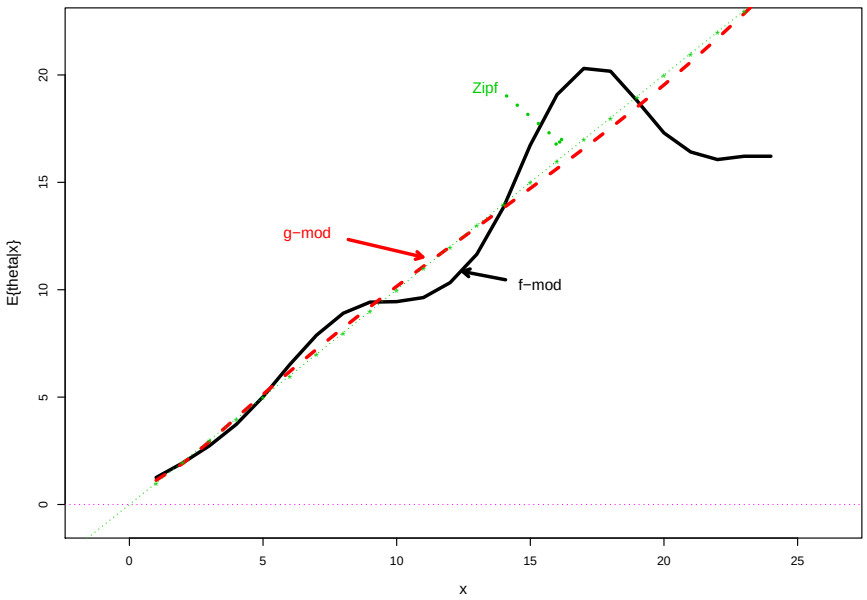

FIG 5. *Solid black curve is Robbins' estimate* (62) *for* $E\{\theta \mid x\}$ *based on natural spline estimate* $\hat{f}(x)$ *in Figure 4; red dashed curve is g-modeling estimate described in the text. It closely follows Zipf's estimate* $E\{\theta \mid x\} = x$.

The solid black curve in Figure 5 shows

$$\hat{e}_g(x) = (x+1)\hat{f}(x+1)/\hat{f}(x) \tag{62}$$

from the Robbins f-modeling estimate in Figure 4. Substituting $f_x = \hat{f}(x)$ and $e_g(x) = \hat{e}_g(x)$ in (61) yields the risk estimate

$$\hat{\mathcal{R}}_g = 6.60 - 2.33 = 4.27. \tag{63}$$

This looks suspect. Robbins' formula has magnified the small bumps seen in Figure 4 into large waves in Figure 5, particularly at the right side where the counts are small. With a sample size of only $N = 501$, it is easy to believe that estimates (62) and (63) are dangerously noisy.

The red dashed curve in Figure 5 is based on *g-modeling*; that is, an estimate of the prior $\hat{g}(\theta)$ has been obtained from the butterfly data by a method described below, directly yielding the posterior expectation

$$e_{\hat{g}}(x) = E_{\hat{g}}\{\theta \mid x\}. \tag{64}$$

Now $\hat{\mathcal{R}}_g = 6.60 - 0.048 = 6.55$ which perhaps seems more reasonable.

Poisson inference problems are often better phrased in terms of the natural parameter

$$\lambda = \log \theta. \tag{65}$$

This is attractive here since the butterfly data is concentrated at small values, where θ itself is a blunt instrument. Tweedie's formulas (24) for the Poisson family are

$$\begin{aligned} e_g(x) &= E\{\lambda \mid x\} = \dot{\text{lgamma}}(x+1) + \dot{l}(x), \\ v_g(x) &= \text{Var}\{\lambda \mid x\} = \ddot{\text{lgamma}}(x+1) + \ddot{l}(x), \end{aligned} \tag{66}$$

the same holding for the truncated Poisson, Remark D. Here lgamma is the log gamma function, the dots indicating first and second derivatives, and $l(x) = \log f(x)$ as before. See Section 2 of Efron (2011).

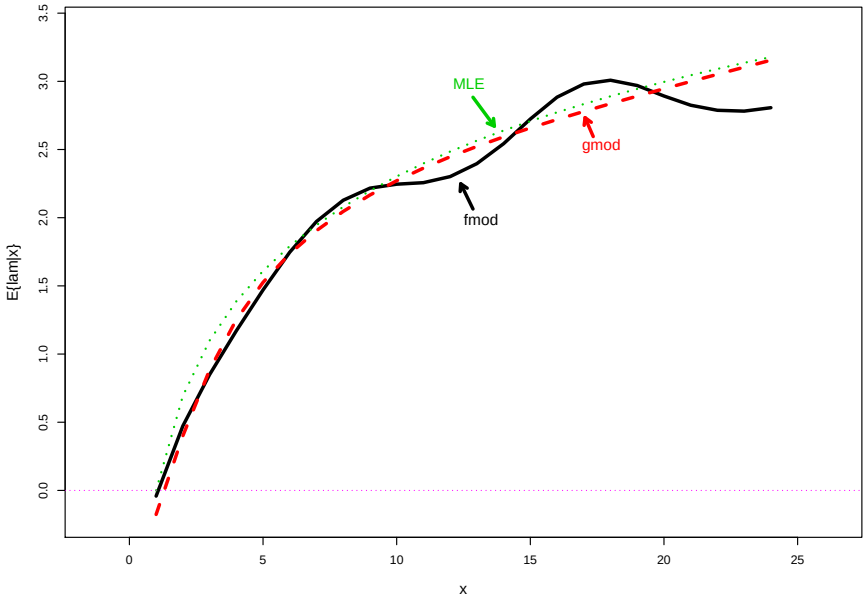

FIG 6. *Estimates of $E\{\lambda \mid x\}$, $\lambda = \log\theta$, butterfly data.* ***Black solid curve:*** *f-modeling estimate, Tweedie's formula using $\hat{f}$ from Figure 4.* ***Red dashed curve:*** *g-modeling estimate as explained in text.* ***Green dotted curve:*** *MLE $(x, \log x)$.*

Figure 6 is the version of Figure 5 that applies to λ rather than θ. The same f-modeling estimate $\hat{f}(x)$ that gave $\hat{e}(x)$ from Robbins' formula (50) now gives the solid black curve "fmod", using (66) with $l(x)$ replaced by $\hat{l}(x) = \log \hat{f}(x)$. Likewise, the same g-modeling estimate $\hat{g}(\theta)$ for the prior density $g(\theta)$ now gives the red dashed curve "gmod" for $\widehat{E}\{\lambda \mid x\}$ using Bayes rule,

$$E_{\hat{g}}\{\lambda \mid x\} = \frac{\int_{\mathcal{T}} \log\theta \hat{g}(\theta) p(x \mid \theta)\, d\theta}{\int_{\mathcal{T}} \hat{g}(\theta) p(x \mid \theta)\, d\theta}. \tag{67}$$

It closely tracks the light dotted green MLE curve $(x, \log x)$, the logarithmic version of Zipf's law.

Table 4 shows estimates of Bayes risk $\mathcal{R}_g$ — or, more directly for the population of 501 species in Table 3, the ASE (5) — and of the empirical Bayes regret (13)

TABLE 4
*Estimates of Bayes risk $\mathcal{R}_g$ and empirical Bayes regret EBregret for $\lambda = \log\theta$, butterfly data. **f-modeling:** $\mathcal{R}_g$ using (18), EBregret from Lemma 4.1 (47). **g-modeling:** $\mathcal{R}_g$ using (18), EBregret as described in Remark F.*

	Bayes risk $\mathcal{R}_g$	EBregret
f-modeling	.316	.002
g-modeling	.334	.018

for λ, for both f- and g-modeling. The risk estimates are not very different, 0.316 versus 0.334, the latter being nearly the same as that for Zipf's rule $\hat{\lambda} = \log x$. Lemma 4.1's estimate EBregret = 0.0024 for f-modeling seems small, but was verified by a bootstrap simulation: 200 multinomial samples $\boldsymbol{y}(j)$ of size $N = 501$ were drawn from $\hat{\boldsymbol{f}}$; $\hat{f}(j)$ and $\hat{e}(j)$, (57) and (62), were calculated; and regret estimated according to the last term in (20). The 200 bootstrap regret estimates averaged 0.0026. Regret associated with g-modeling was estimated by a method described in Section 8, Remark F.

The basic idea of g-modeling (Efron, 2016) is simple: the prior density $g(\theta)$ is modeled as a low-dimensional exponential family, for example,

$$\log g_\beta(\theta) = \sum_{j=0}^{J} \beta_j \theta^j; \tag{68}$$

$g_\beta(\cdot)$ induces a marginal density $f_\beta(x)$ as in (3); finally, $\hat{g} = g_{\hat{\beta}}(\cdot)$ is found by numerical maximization of the log likelihood,

$$\hat{\beta} = \arg\max_\beta \left\{ \sum_{i=1}^{N} \log f_\beta(x_i) \right\}. \tag{69}$$

Some details appear in Remark F of Section 8.

TABLE 5
The g-modeling estimate of $E\{\lambda \mid x\}$ in Figure 6. Comparison of the posterior standard deviation of λ given x with the frequentist root mean square error of $E\{\lambda \mid x\}$.

x	$E\{\lambda \mid x\}$	$\mathrm{sd}\{\lambda \mid x\}$	Freq RMSE
2	0.40	.731	.078
6	1.74	.432	.033
10	2.27	.313	.024
14	2.59	.260	.019
18	2.84	.237	.024
22	3.05	.229	.058

For the butterfly data, $g(\theta)$ was assumed to follow a natural spline with five degrees of freedom; this is a version of (68) with the powers θ^j replaced by a different set of basis polynomials, *B-splines* (Hastie, Tibshirani and Friedman, 2009, Chap. 5). Table 5 shows the resulting estimate of the posterior mean and standard deviation of λ given x.

The final column gives the *frequentist* root mean square errors (RMSEs) of $\widehat{E}\{\lambda \mid x\}$, the red dashed curve in Figure 6, which are seen to be rather small.

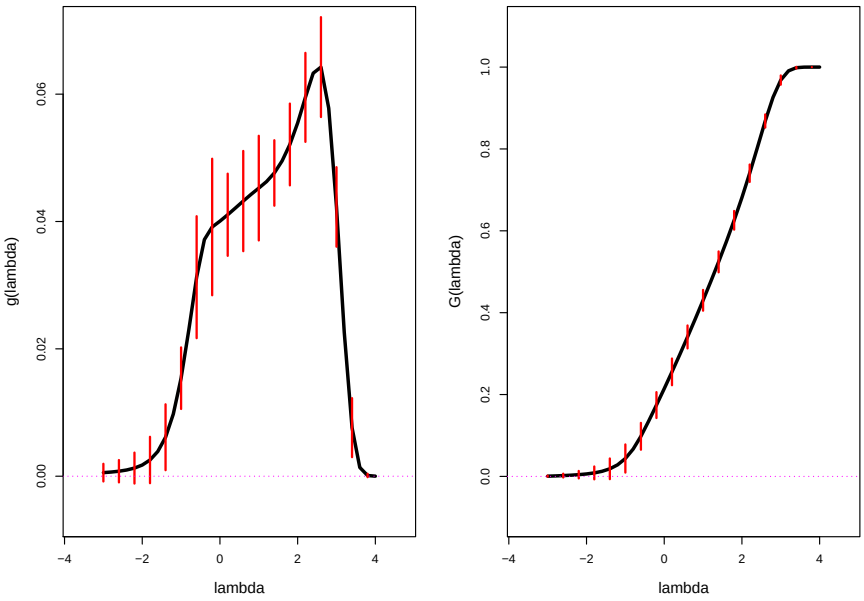

FIG 7. ***Left panel:*** *Estimated prior density for* $\lambda = \log\theta$*, butterfly data, based on natural spline model with 5 degrees of freedom for* $g(\theta)$*.* ***Right panel:*** *Corresponding estimate of cdf. Red bars indicate* ± 1 *root mean square frequentist error; see Remark F.*

Empirical Bayes estimation, more than full Bayes, encourages frequentist calculations of accuracy. Remark F reviews the RMSE calculations.

The estimate of prior density $g(\theta)$ obtained by maximum likelihood in the natural spline model is graphed in the left panel of Figure 7. Actually, $\hat{g}(\theta)$ has been transformed to a density $\tilde{g}(\lambda)$ for $\lambda = \log\theta$, i.e., $\tilde{g}(\lambda) = \hat{g}(\theta)\theta$, to avoid the pile-up of $g(\theta)$ near $\theta = 0$. The right panel shows the estimated cdf $\widehat{G}(\theta) = \int_0^\theta \hat{g}(t)\,dt$, again plotted versus λ. Speaking loosely, λ is close to uniform between -1 and 3.

The red vertical bars in Figure 7 indicate $\pm$ one frequentist root mean square error. We see that the cdf is estimated more accurately than the density itself. Empirical Bayesian estimation of quantities beyond the scope of f-modeling are permitted by g-modeling, for instance $\widehat{\Pr}\{\theta \le 1 \mid x = 3\}$ (calculated to be 0.126 here).

Empirical Bayes can be said to begin with Corbet's question to Fisher: "How many new species can I expect to find in one more year of trapping?" It can be shown that the expected number of new species in t years of additional trapping, say new(t), is

$$E\{\text{new}(t)\} = N\int_{\mathcal{T}} e^{-\theta}\frac{1-e^{-\theta t/2}}{1-e^{-\theta}}\,g(\theta)\,d\theta. \tag{70}$$

See Remark G. The solid curve in Figure 8 shows the g-modeling values of new(t), with frequentist standard deviation indicated by vertical bars. At year $t = 1$ we get

$$E\{\text{new}(1)\} = 47.6 \pm 4.4. \tag{71}$$

Good and Toulmin's (1956) nonparametric f-modeling estimate, indicated by red dots in Figure 8, gave 45.2 ± 9.3. See Section 11.5 of Efron (2010).

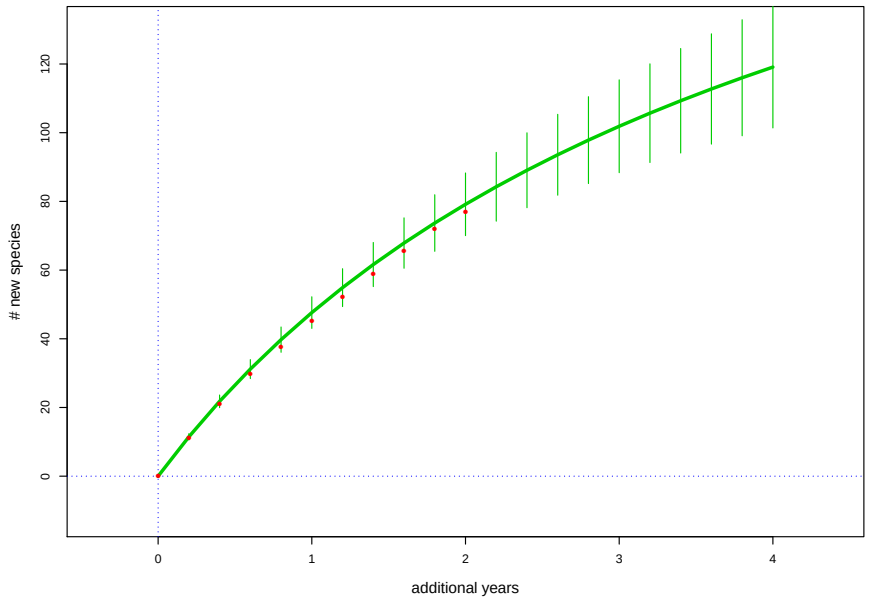

FIG 8. *Estimated number of new butterfly species captured in t additional years of trapping, obtained by substituting g-modeling estimate $\hat{g}(\theta)$ in* (70). *Red dots show Good and Toulmin's nonparametric f-modeling estimate; green vertical bars indicate* ± 1 *frequentist RMSE for the solid curve.*

Once again, the assumption $\theta_i \overset{\text{ind}}{\sim} g(\theta)$ (2) plays only a motivational role here; $\hat{g}(\theta)$ in Figure 7 estimates $\bar{g}(\theta)$, the empirical density of $\boldsymbol{\theta}_{\text{ord}}$ (7), regardless of $\boldsymbol{\theta}_{\text{ord}}$'s provenance. We don't have a butterfly oracle for guidance but Table 4 says we hardly need one. The more-Bayesian side of empirical Bayes analysis shows itself in the next section, where we consider posterior inferences for individual parameters θ_i.

6. FINITE BAYES INFERENCE

We return to empirical Bayes model (1)–(2),

$$\theta_i \overset{\text{ind}}{\sim} g(\theta) \quad \text{and} \quad x_i \overset{\text{ind}}{\sim} \mathcal{N}(\theta_i, 1) \tag{72}$$

for $i = 1, 2, \ldots, N$, with $g(\cdot)$ unknown and the θ_i unobserved. One more x has been observed, independent of the N other observations, say

$$x_0 \sim \mathcal{N}(\theta_0, 1), \tag{73}$$

withe the unobserved θ_0 drawn independently from $g(\cdot)$. Our goal is to assess the posterior distribution of θ_0 given x_0 and $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$. Unlike ASE in Section 2, now we are specifically interested in θ_0, not some omnibus loss function over all the θ_i's.

An example appears in Figure 9: $x_0 = 5$, while of the $N = 50$ others $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$, 47 are less than 5. What can we say about θ_0? This can be called the *finite Bayes inference* problem. If N were infinity we could deconvolute $\boldsymbol{x}$ to learn $g(\theta)$ exactly, and then use Bayes rule to calculate $g(\theta_0 \mid x_0)$–which is to say that standard Bayes is finite Bayes with $N = \infty$.

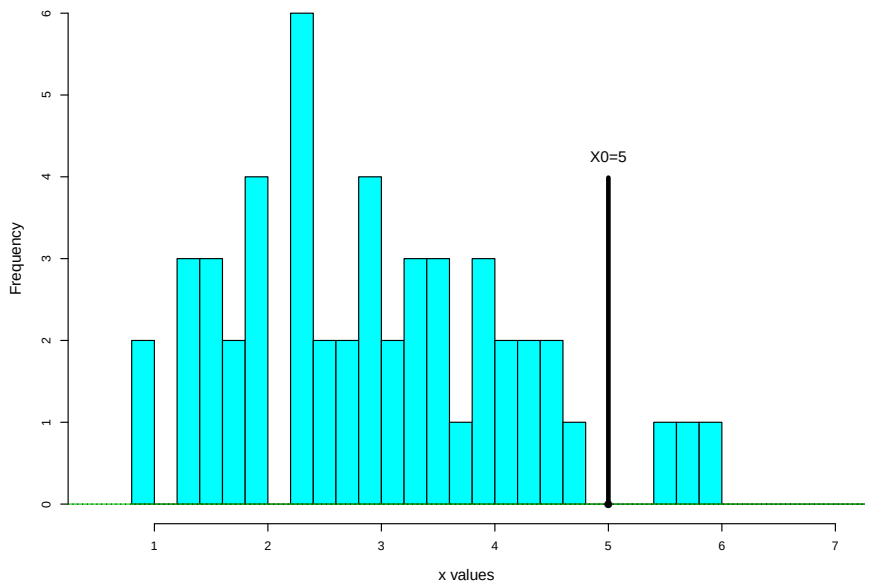

FIG 9. *Finite Bayes inference. We observe* $x_0 = 5$ *and* $N = 50$ *other values* $x_1, x_2, \ldots, x_N$ *indicated by the histogram. All 51* x_i*'s are obtained independently as in* (72)*. What can we say about* θ_0*, the parameter that produced* x_0*?*

A fully Bayesian approach to the finite Bayes problem would begin by putting a hyperprior $h(g)$ on the choice of $g(\cdot)$. This is the *Bayes empirical Bayes* approach of Deely and Lindley (1981). Choosing $h(\cdot)$ is an uncertain task, however, and having done so it still can be difficult to compute the resulting posterior distribution for θ_0. Instead, we will employ empirical Bayes g-modeling estimates $\hat{g}(\cdot)$, g-modeling being necessary here for the calculation of $\hat{g}(\theta_0 \mid x_0)$. Now the assumption that all the θ's are generated from $\theta \sim g(\cdot)$ is crucial. It is what makes the "sibling" observations $x_1, x_2, \ldots, x_N$ relevant to the inference of θ_0.

Morris (1983) considered the question of setting accurate empirical Bayes confidence intervals in the case where the prior density is normal, the James–Stein case. In the simplest situation we have

$$\theta_i \overset{\text{ind}}{\sim} \mathcal{N}(0, A), \qquad x_i \mid \theta_i \overset{\text{ind}}{\sim} \mathcal{N}(\theta_i, 1), \tag{74}$$

for $i = 1, 2, \ldots, N$, A unknown, so that

$$\theta_i \mid x_i \sim \mathcal{N}(Bx_i, B) \qquad [B = A/(A+1)]\,. \tag{75}$$

The James–Stein rule $\hat{\theta}_i = \widehat{B}x_i$ substitutes the unbiased estimate

$$\widehat{B} = 1 - (N-2) \Big/ \sum_{i=1}^{N} x_i^2 \tag{76}$$

for B. Looking at (75), this suggests

$$\widehat{B}x_i \pm z^{(1-\alpha/2)}\widehat{B}^{1/2} \tag{77}$$

as a level $1-\alpha$ posterior interval for θ_i given x_i, where $z^{(\alpha)}$ is the standard normal quantile $\Phi^{-1}(\alpha)$, e.g., $z^{(0.95)} = 1.96$.

The trouble, as Morris points out, is that (77) doesn't take into account the variability of $\widehat{B}$ as an estimate of B. A wider interval,

$$\widehat{B}x_i \pm z^{(1-\alpha/2)} \left\{ \widehat{B} + \frac{2}{N-2} \left[x_i \left(1 - \widehat{B} \right) \right]^2 \right\}^{1/2}, \tag{78}$$

is necessary to give more accurate $1-\alpha$ coverage. With $N = 20$, $\widehat{B} = 1/2$, and $x_i = 3$, for example, (78) is 22% wider than (77). Interval (78) approximates what we would get from a full Bayesian analysis of (74) that began with an uninformative hyperprior on A.

Morris' intervals are based on the assumption of a Gaussian prior. Here we will discuss g-modeling approaches to more general finite Bayes inference problems, substituting computer power for mathematical analysis in going from the equivalent of (77) to (78).

The finite Bayes computations of this section proceed in five steps:

1. Data set $\boldsymbol{x}$ gives an estimated prior density $\hat{g}(\theta)$ by g-modeling.
2. The estimated marginal density $\hat{f}(x) = \int_{\mathcal{T}} \hat{g}(\theta) p(x \mid \theta)\, d\theta$ is computed.
3. Parametric bootstrap data sets $\boldsymbol{x}^* = (x_1^*, x_2^*, \ldots, x_N^*)$ are drawn fronm $\hat{f}(\cdot)$,

$$x_i^* \overset{\text{ind}}{\sim} \hat{f}(\cdot) \qquad \text{for } i = 1, 2, \ldots, N. \tag{79}$$

4. Data set $\boldsymbol{x}^*$ gives $\hat{g}^*(\theta)$ using the same g-model as in Step 1.
5. Some large number B of bootstrap priors $\hat{g}^*(\cdot)$ are averaged to give a *corrected* prior,

$$\tilde{g}(\theta) = \frac{1}{B} \sum_{j=1}^{B} \hat{g}^{*j}(\theta). \tag{80}$$

The idea here, taken from Laird and Louis (1987), is that the bootstrap distribution of $\hat{g}^*(\theta)$ mimics the posterior variability of $g(\theta)$ given $\boldsymbol{x}$ in a full Bayesian analysis that began with an uninformative hyperprior $h(g)$. If so, the corrected posterior density

$$\tilde{g}(\theta_0 \mid x_0) = \tilde{g}(\theta_0) p(x_0 \mid \theta_0) / \tilde{f}(x_0) \tag{81}$$

— here $\tilde{f}(x_0) = \int_{\mathcal{T}} p(x_0 \mid \theta) \tilde{g}(\theta)\, d\theta$, with $p(x \mid \theta) = \phi(x - \theta)$ — approximates $g(\theta_0 \mid x_0)$ from a full Bayesian analysis.

The solid black curve in Figure 10 graphs $\tilde{g}(\theta_0 \mid x_0 = 5)$ from the 50 observations in Figure 9. It assumed model (72), and was computed using the five-step algorithm; the g-model was a natural spline with five degrees of freedom, with $B = 1000$ in (80). The green dotted curve is $\tilde{g}(\theta)$, while the red dashed curve is the likelihood function $\phi(\theta_0 - x_0)$ for θ_0 given just $x_0 = 5$–that is, ignoring the 50 sibling observations. *Not* ignoring them has a powerful effect on our beliefs concerning θ_0: $\tilde{g}(\theta_0 \mid x_0)$ has its maximum at $\theta_0 = 3.8$, compared to the MLE 5, and puts only 18% of its posterior probability above 5.

The *gamnormal* example (48) comprises $N = 3200$ values θ_i obtained from a $\text{Gamma}_9/3$ distribution (mean = 3 and variance = 1) and 3200 corresponding observations $x_i \overset{\text{ind}}{\sim} \mathcal{N}(\theta_i, 1)$. The first 50 x_i's are those in the histogram of Figure 9. A light black beaded curve in Figure 10 traces the true posterior density

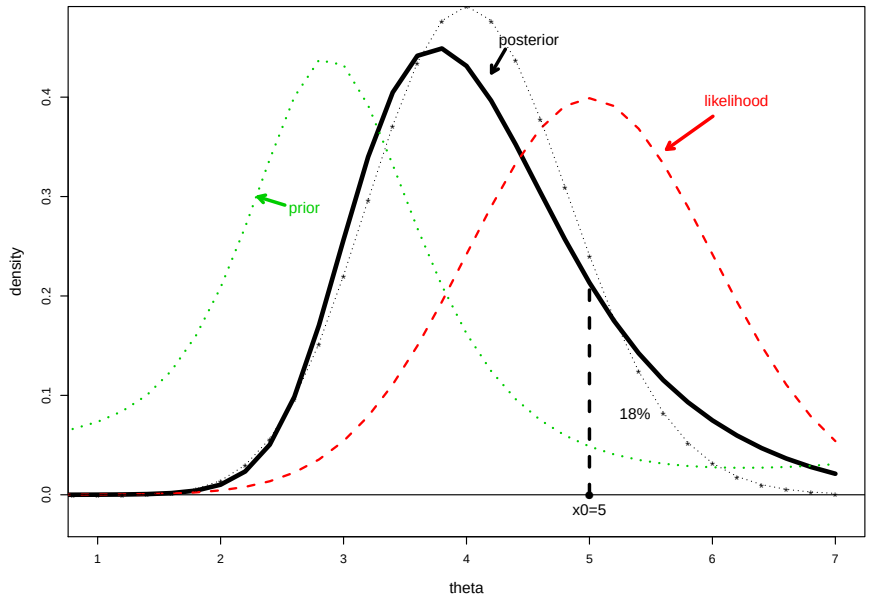

FIG 10. ***Solid black curve:*** *Corrected posterior density* $\tilde{g}(\theta_0 \mid x_0 = 5)$ *from the 50 observations* x_i *in Figure 9, using model ns(*θ*,df* $= 5$*).* ***Green dotted curve:*** *Estimated prior* $\tilde{g}(\theta)$ (80) $B = 1000$. ***Red dashed curve:*** *Likelihood* $\phi(\theta - x_0)$. ***Light beaded curve:*** *True posterior density* $g(\theta_0 \mid x_0)$ *based on* $\text{Gamma}_9/3$ *prior.*

TABLE 6

Ratio of spreads of corrected densities $\tilde{g}(\theta)$ (80) *compared to uncorrected g-model estimates* $\hat{g}(\theta)$*, assuming natural spline model with df* $= 5$*. Data* $\boldsymbol{x}$ *is first* N *observations of the 3200 gamnormal draws. Each* $\tilde{g}(\cdot)$ *based on* $B = 1000$ *bootstrap replications. "Spread" is distance between* α*th and* $(1-\alpha)$*th quantiles, averaged over* $\alpha = 0.90, 0.80, 0.70, 0.60$.

N	15	25	50	100	200	400	800	1600	3200
ratio	1.55	1.45	1.27	1.11	1.06	1.04	.96	.97	.96

$g(\theta_0 \mid x_0 = 5)$ based on the $\text{Gamma}_9/3$ prior; $\tilde{g}(\theta_0 \mid x_0)$ is seen to be reasonably accurate considering its basis of only 50 siblings.

Correction (80) is impactful in this case, both $\tilde{g}(\theta)$ and $\tilde{g}(\theta_0 \mid x_0 = 5)$ being more than 25% wider than the uncorrected versions. Increasing the number N of sibling observations, from 50 to $100, 200, \ldots, 3200$, quickly decreases correction effects, as seen in Table 6. Even for $N = 50$, $\tilde{g}(\theta_0 \mid x_0)$ was only a modest improvement over the uncorrected $\hat{g}(\theta_0 \mid x_0)$ as far as comparisons with the true $g(\theta_0 \mid x_0)$ go.

Correction method (80) has its critics — Carlin and Gelfand (1991) and Section 5 of Efron (1996) — who provide more accurate but also more involved bootstrap algorithms. Applied to the Morris Gaussian prior situation (74), (80) gives corrections similar to (78), e.g., 27% dilation compared to 22% in the example following (78). Laird and Louis (1987) provide some favorable simulation results. As Table 6 suggests, correction effects are likely to be small when N is in the thousands range. In any case, the bootstrap replications $\hat{g}^{*j}(\cdot)$ (80) can also be used to assess frequentist standard errors — of $\hat{g}(\theta \mid x)$, $\widehat{E}\{\theta \mid x\}$, etc. — which will be the same as the Laird–Louis assessments of Bayesian accuracy.

How many sibling observations are enough? An answer must depend on the

shape of the true prior density $g(\theta)$ and the assumptions of the g-modeling procedure. In the gamnormal example, employing a natural spline model with df $= 5$, useful results were obtained for N as small as 15.

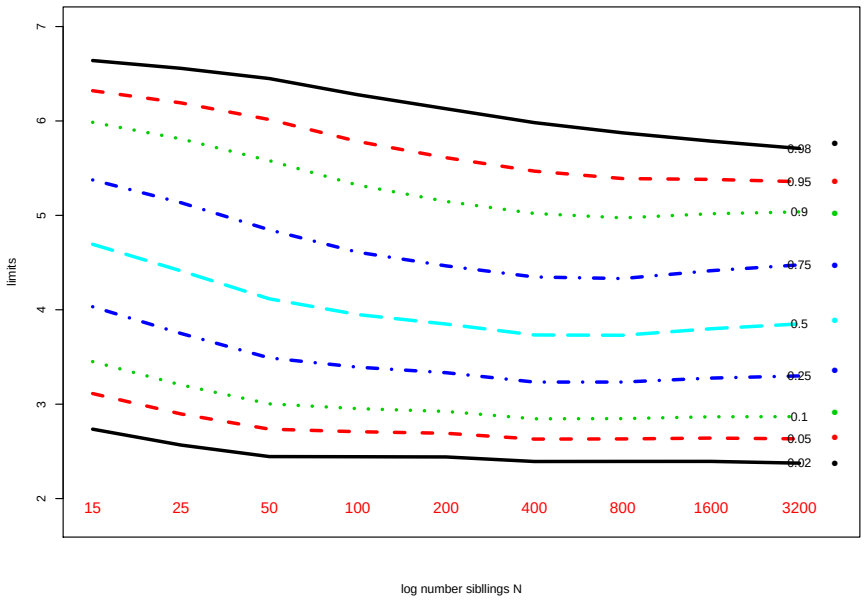

FIG 11. *Smoothed posterior percentiles* $2\%, 5\%, \ldots, 98\%$ *of corrected estimates* $\tilde{g}(\theta_0 \mid x_0 = 5)$ *based on first* N *siblings, gamnormal example, as* N *increases. Points at extreme right are percentiles of true posterior* $g(\theta_0 \mid x_0 = 5)$ *for* $\theta \sim \text{Gamma}_9/3$. *See Remark H.*

This doesn't mean that larger values of N are pointless. Figure 11 graphs the percentiles of the corrected posterior densities $\tilde{g}(\theta_0 \mid x_0 = 5)$ as N increases from 15 to 3200 in the gamnormal example; $B = 1000$ bootstraps for each N. (There has been some smoothing; see Remark H.) What is perhaps surprising is that some "learning" is going on even for large N, as seen most vividly in the decline of the 0.98 percentile curve.

Points at the extreme right of Figure 11 show percentiles of the true posterior density $g(\theta_0 \mid x_0 = 5)$. These are not quite the same as what we would get by extending the figure's range toward $N = \infty$. The class of prior densities obtainable from natural spline models with five degrees of freedom does not include the $\text{Gamma}_9/3$ density, causing a small amount of *modeling bias*.

Nonparametric g-modeling is an appealing remedy for modeling bias: in empirical Bayes situations such as (49)–(72), we find the prior distribution that maximizes the likelihood of the observed data $\boldsymbol{x}$ without restrictions on the form of $g(\cdot)$. Impressive theoretical work on nonparametric maximum likelihood (NPMLE) solutions (Kiefer and Wolfowitz, 1956; Laird, 1978) still left the problem computationally forbidding. Progress in convex optimization (Gu and Koenker, 2016) has now crossed that river. Extensive theoretical and computational calculations in Jiang and Zhang (2009) demonstrate excellent performance for NPMLE methods, for model (72), in terms of the ASE criterion (5).

Application of g-modeling to the full gamnormal data set, $N = 3200$, were carried out using natural spline models ns(θ, df) with df $= 5$, 20, and 80. The last of these approximate NPMLE. Figure 12 shows the resulting uncorrected

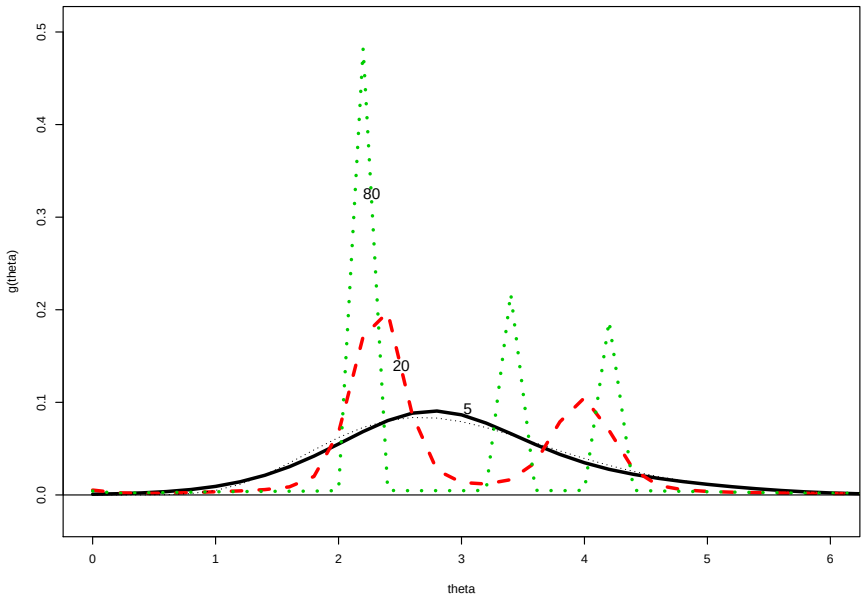

FIG 12. *Prior density estimates $\hat{g}(\theta)$ based on all $N = 3200$ gamnormal observations. For natural spline g-models and degrees of freedom 5, 20, and 80. Light black dashed curve is the true prior density* $\text{Gamma}_9/3$.

estimates $\hat{g}(\theta)$. For df $= 80$, $\hat{g}(\theta)$ is almost a discrete distribution supported on three points, in agreement with Laird's characterization of NPMLE solutions. The three $\hat{g}(\cdot)$ estimates had EBregret 0.008, 0.009, and 0.006, respectively. The oracle ASE (5) was 0.489, making empirical Bayes regret negligible in this case.

In the finite Bayes setup, the sibling observations x_i are related to the object of interest θ_0 through the Bayesian relationship $\theta_i \sim g(\cdot)$ for $i = 0, 1, 2, \ldots, N$. Suppose instead the relationship is through a regression model

$$\theta_i = c_i'\beta \qquad (i = 0, 1, 2, \ldots, N), \tag{82}$$

where the c_i are known covariate vectors and β is an unknown parameter vector. Under mild conditions, as $N \to \infty$ the MLE $\hat{\theta}$ converges in distribution to the true value θ_0. This isn't the case for the finite Bayes situation, where the best we can hope for is convergence to the true posterior density $g(\theta_0 \mid x_0)$. In this sense, sibling observations are weaker than regression observations $x_i = c_i'\beta + \epsilon_i$ but, as Figure 10 shows, they can still have a powerful effect on our beliefs about θ_0.

The NPMLE approach is less appropriate for finite Bayes inference problems. Applied to the situation in Figure 9, it would give a posterior estimate of $g(\theta_0 \mid x_0)$ supported on just a few discrete points. A "smooth" model for $g(\cdot)$, such as ns(θ, df $= 5$), is a bet on its smoothness. Betting isn't necessary for omnibus criteria like ASE but becomes crucial for finite Bayes calculations.

7. RELEVANCE

We return to the question of relevance raised at the end of Section 2: which other cases are relevant to our beliefs concerning a particular parameter θ_0?

Questions of relevance can be especially pressing when the individual observations are accompanied by covariate information. Such a situation is illustrated in

Figure 13: 12 children, six dyslexic and six normal controls have received a DTI (diffusion tensor imaging) scan, measuring fluid flow at $N = 15,443$ brain locations or *voxels*. Each voxel provided a two-sample statistic z comparing dyslexics with normal controls, with

$$z_i \sim \mathcal{N}(\delta_i, 1), \qquad i = 1, 2, \ldots, N = 15,443, \tag{83}$$

to a good approximation; δ_i is the effect size for voxel$_i$, and of course the investigators were interested in voxels having δ_i much different than zero. (See Section 15.6 of Efron and Hastie, 2016.)

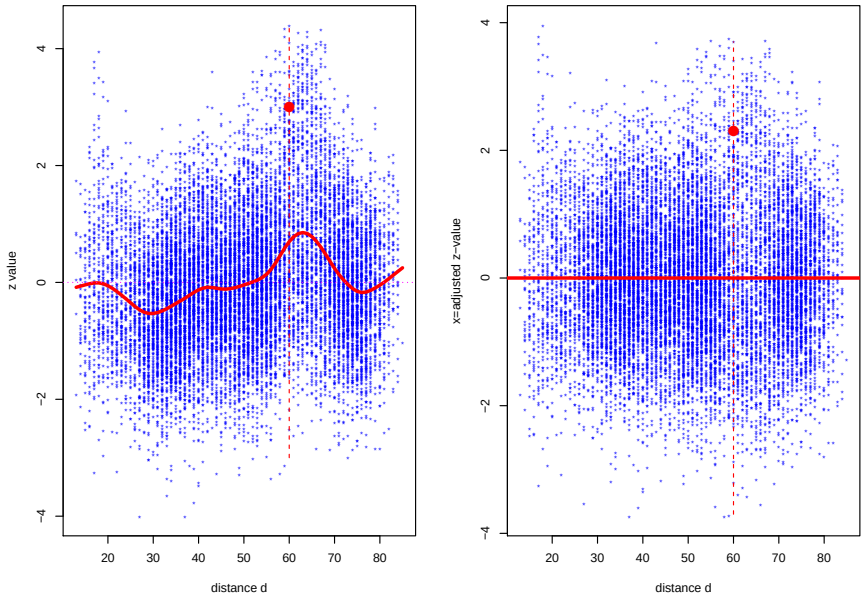

FIG 13. *DTI data.* ***Left panel:*** *z-values z_i plotted vertically versus voxel distance from back of the brain d_i; red curve $c(d)$ is smoothing spline, $df = 13$; large red dot $(d_0, z_0) = (60, 3.0)$. What is the posterior distribution of δ_0, the expectation of z_0?* ***Right panel:*** *Vertical axis is $x_i = z_i - c(d_i)$; now the red dot at $d_0 = 60$, $x_0 = 3.0 - 0.7 = 2.3$. The expectation of x_0 is $\theta_0 = \delta_0 - 0.7$.*

The left panel of Figure 13 plots the z_i vertically versus d_i, the voxel distance from the back of the brain. The large red dot indicates "voxel$_0$", a location where effect size δ_0 is of particular interest. It has coordinates

$$(d_0, z_0) = (60, 3.0). \tag{84}$$

What can we say about δ_0, based on (84) and the 15442 other (d_i, z_i) observations?

The *distance* covariate induces substantial effects, raising or lowering the entire distribution of z-values for varying values of d. A smoothing spline with 13 degrees of freedom, $c(d)$, fit to z_i as a function of d_i, appears as the solid red curve in the left panel. Subtracting $c(d)$ from the observations z_i yields standardized values x_i,

$$x_i = z_i - c(d_i). \tag{85}$$

In what follows we will analyze the model

$$x_i \sim \mathcal{N}(\theta_i, 1) \qquad (\theta_i = \delta_i - c(d_i)); \tag{86}$$

see Remark I. Since $c(d_0 = 60) = 0.70$, we have

$$\theta_0 = \delta_0 - 0.70 \tag{87}$$

as the parameter of interest, with the red dot corresponding to

$$x_0 = z_0 - 0.70 = 2.30. \tag{88}$$

The adjusted points (d_i, x_i) plotted in the right panel seem better behaved, but still with some heterogeneity visible as a function of d. We wish to calculate a finite Bayes posterior distribution for θ_0. First though, we have to decide which of the $N - 1$ other points x_i are legitimate siblings for x_0. All of them? Only those with $50 \le d \le 70$? Only those with $d = 60$?

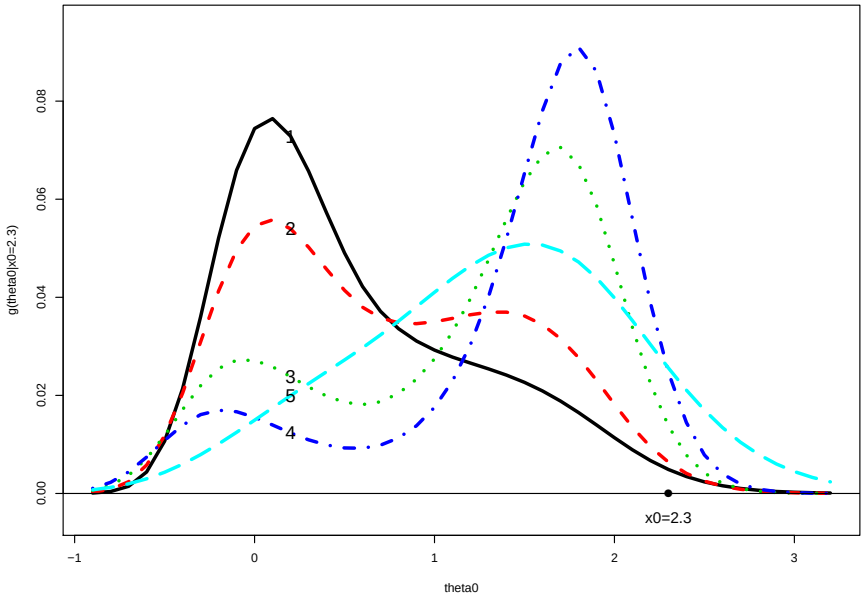

FIG 14. *g-modeling estimates of $\hat{g}(\theta_0)$, the finite Bayes posterior density for θ_0 (87) for the five choices of sibling set shown in Table 7.*

Uncorrected g-modeling estimates $\hat{g}(\theta_0 \mid x_0 = 2.3)$ were computed using a natural spline model with five degrees of freedom, and with five different choices of the sibling set as described in Table 7. The resulting estimated posterior densities $\hat{g}(\theta_0 \mid x_0 = 2.3)$ appear in Figure 14, numbered as in that list. In this case, choice (1), using all 15,442 others as siblings, moves the estimated conditional distribution of θ_0 given $x_0 = 2.3$ to the left, compared with the more restrictive choices (2) through (5). Table 8 provides some numerical comparisons: $E\{\theta_0 \mid x_0\}$ increases from 0.557 for choice (1) to 1.32 for choice (5), while the posterior probability of θ_0 exceeding 2.3 goes from 0.036 to 0.197.

Restricting the sibling set in the name of greater relevance can potentially destabilize the estimated posterior density $\hat{g}(\theta_0 \mid x_0)$. This is seen in Table 8, where decreasing the sample size N increases the frequentist standard error of the estimated posterior expectation $E\{\theta_0 \mid x_0\}$, most noticeably for the smallest set, $N = 186$.

None of this is very reassuring. Adjusting for covariates — going from the left panel to the right in Figure 13 — is helpful in strengthening relevance, but at

TABLE 7
Five choices of the sibling set for x_0, from $N - 1 = 15442$ other points (d_i, z_i), DTI data.

(1) All 15442 others.
(2) Those with $40 \le d \le 80$.
(3) Those with $50 \le d \le 70$.
(4) Those with $55 \le d \le 65$.
(5) Those with $d = 60$.

TABLE 8
***Top 2 rows:** Finite Bayes posterior expectation and standard deviation of θ_0 (87) for the 5 decreasing choices of sibling set shown in Table 7. **3rd row:** Frequentist standard error of the top-row estimate; see Remark F. **4th row:** Estimated posterior probability that θ_0 exceeds 2.3. **5th row:** Sample sizes N of the five choices in Table 7.*

	(1) all	(2) 40–80	(3) 50–70	(4) 55–65	(5) 60
$E\{\theta_0 \mid x_0 = 2.3\}$	.557	.735	1.15	1.40	1.32
$\mathrm{sd}\{\theta_0 \mid x_0 = 2.3\}$	.692	.747	.796	.769	.779
frequentist $\mathrm{sd}(E)$	.043	.048	.051	.060	.146
$\Pr\{\theta_0 > 2.3 \mid x_0 = 2.3\}$	.036	.048	.109	.186	.197
N	15443	10462	5249	2401	186

least in this example it is not a cure. At some level, the choice of relevant siblings is a scientific question rather than a purely statistical one. Perhaps we can expect the neuroscientists who provided the DTI data to say what relevance means here; perhaps not. In Bayes (as opposed to empirical Bayes) applications, the assertion of a prior density $g(\theta)$ amounts to a statement of an infinite catalog of perfectly relevant siblings. Empirical Bayes applications like that in Figure 13 illustrate the sometimes heroic nature of such statements.

There are situations where purely statistical evidence might cast doubt on relevance, for example in Figure 9, where 47 of the 50 putative siblings lie to the left of the index case $x_0 = 5$. A procedure for allowing discordant values of x_0 to "opt out" of an empirical Bayes analysis is described in Section 4 of Efron (1996): we assume that the prior density $g(\theta)$ is a mixture of two components, a main one $g_A(\theta)$ and a much broader opt-out alternative $g_B(\theta)$,

$$g(\theta) = h_A g_A(\theta) + h_B g_B(\theta), \tag{89}$$

h_A and $h_B = 1 - h_A$ being the respective hyperprior probabilities. This gives marginal density $f(\theta)$ (3),

$$f(x) = h_A f_A(x) + h_B f_B(x), \tag{90}$$

with $f_A(x)$ and $f_B(x)$ the respective marginals. In what follows, we will set

$$f_B(x) = f_0(x)/c, \tag{91}$$

where $f_0(x)$ is a given, possibly improper, density function and c is a constant to be determined.

Bayes rule yields posterior probabilities $h_A(x)$ and $h_B(x)$ given x, with

$$\frac{h_A(x)}{h_B(x)} = \frac{h_A}{h_B}\,\frac{f_A(x)}{f_B(x)} = c\,\frac{h_A}{h_B}\,\frac{f_A(x)}{f_0(x)}, \tag{92}$$

or equivalently,

$$h_B(x) = \left[1 + c\,\frac{h_A}{h_B}\,\frac{f_A(x)}{f_0(x)}\right]^{-1}. \tag{93}$$

Since it is always true that $h_B = \int_{\mathcal{X}} h_B(x) f(x)\,dx$, we get the identity

$$h_B = \int_{\mathcal{X}} f(x)\left[1 + c\,\frac{h_A}{h_B}\,\frac{f_A(x)}{f_0(x)}\right]^{-1} dx. \tag{94}$$

This determines c, and also $h_B(x)$ (93).

The opt-out analysis was applied to the situation in Figure 10, where the observed data is $\boldsymbol{x} = (x_1, x_2, \ldots, x_{50})$, with the following specifications:

- $h_A = 0.95$ and $h_B = 0.05$.
- $f_0(x) = 1$ for all x.
- $f_A(x)$ equal $\tilde{f}(x)$ the marginal density obtained from $\tilde{g}(\cdot)$, the green dotted curve. (See Remark J.)
- The expectation with respect to the marginal $f(x)$ in (94) replaced by

$$\frac{1}{50}\sum_{i=1}^{50}\left[1 + c\,\frac{h_A}{h_B}\,\frac{f_A(x_i)}{f_0(x_i)}\right]^{-1}. \tag{95}$$

Then (93) gave $h_B(5) = 0.088$ as the estimated probability that $x_0 = 5$ is not from the same model (72) that produced $x_1, x_2, \ldots, x_{50}$. In terms of Figure 10, the posterior distribution of θ_0 given $x_0 = 5$ and $\boldsymbol{x}$ can be thought of as a mixture that is 91.2% of the solid black posterior curve and 8.8% of the red dashed likelihood; the posterior probability that θ_0 exceeds 5 rises from 18% to 21%.

8. REMARKS

This section expands on some of the points raised elsewhere in this article.

A. Tweedie's formula (24) Differentiating $p_\theta(x)$ (23) with respect to x,

$$\dot{p}(x \mid \theta) = \frac{\partial p(x \mid \theta)}{\partial x} = \left(\theta - \frac{\partial}{\partial x}\log p_0(x)\right) p_\theta(x), \tag{96}$$

gives the derivative of the marginal density $f(x) = \int_{\mathcal{T}} p(x \mid \theta) g(\theta)\,d\theta$,

$$\dot{f}(x) = \int_{\mathcal{T}} \left(\theta - \frac{\partial}{\partial x}\log p_0(x)\right) p(x \mid \theta) g(\theta)\; d\theta. \tag{97}$$

Therefore $l(x) = \log f(x)$ has

$$\begin{aligned} \dot{l}(x) &= \frac{\int_{\mathcal{T}} \left(\theta - \frac{\partial}{\partial x}\log p_0(x)\right) p(x \mid \theta) g(\theta)\; d\theta}{\int_{\mathcal{T}} p(x \mid \theta) g(\theta)\; d\theta} \\ &= E\{\theta \mid x\} - \frac{\partial}{\partial x}\log p_0(x), \end{aligned} \tag{98}$$

or

$$E\{\theta \mid x\} = \dot{l}(x) - \dot{l}_0(x), \tag{99}$$

Tweedie's formula (24). This demonstrates the necessary fact that Tweedie's formula gives the same value of $E\{\theta \mid x\}$ as direct application of Bayes rule.

B. Tweedie's formula for $x \sim \mathcal{N}(\theta, \sigma^2)$ With σ^2 known, (26) becomes

$$e_g(x) = x + \sigma^2 \dot{l}(x) \quad \text{and} \quad v_g(x) = \sigma^2 \left(1 + \sigma^2 \ddot{l}(x)\right). \tag{100}$$

C. Lemma 4.1 Formula (47) will be justified here using plug-in substitutions ($\hat{\boldsymbol{f}}$ for $\boldsymbol{f}$, etc.) and discrete notation as in (34). Letting $\hat{\boldsymbol{l}} = \log \hat{\boldsymbol{f}} = (\log \hat{f}_1, \log \hat{f}_2, \ldots, \log \hat{f}_K)$, the $N \times N$ derivative matrix of $\hat{\boldsymbol{l}}$ with respect to $\boldsymbol{y}$ in (37) can be shown to be

$$\frac{d\hat{\boldsymbol{l}}}{d\boldsymbol{y}} = \boldsymbol{M}\widehat{\boldsymbol{G}}^{-1}\boldsymbol{M}' \qquad \left(\widehat{\boldsymbol{G}} = \boldsymbol{M}' \operatorname{diag}\left(N \cdot \hat{\boldsymbol{f}}\right) \boldsymbol{M}\right), \tag{101}$$

$\operatorname{diag}(N \cdot \hat{\boldsymbol{f}})$ the diagonal matrix with entries $N\hat{\boldsymbol{f}}_k$. The vector of derivatives $\hat{\boldsymbol{l}}^{\cdot} = \boldsymbol{D}\hat{\boldsymbol{l}}$ then has derivative matrix

$$\frac{d\hat{\boldsymbol{l}}^{\cdot}}{d\boldsymbol{y}} = \dot{\boldsymbol{M}}\widehat{\boldsymbol{G}}^{-1}\dot{\boldsymbol{M}}', \tag{102}$$

as in (45). Since $\boldsymbol{y}$ has covariance matrix $\operatorname{diag}(N \cdot \boldsymbol{f})$, estimated as $\widehat{\operatorname{cov}}(\boldsymbol{y}) = \operatorname{diag}(N \cdot \hat{\boldsymbol{f}})$ (102) yields the delta method estimate of $\operatorname{cov}(\hat{\boldsymbol{l}}^{\cdot})$,

$$\widehat{\operatorname{cov}}\left(\hat{\boldsymbol{l}}^{\cdot}\right) = \dot{\boldsymbol{M}}\widehat{\boldsymbol{G}}^{-1}\dot{\boldsymbol{M}}' \operatorname{diag}\left(N \cdot \hat{\boldsymbol{f}}\right) \dot{\boldsymbol{M}}\widehat{\boldsymbol{G}}^{-1}\dot{\boldsymbol{M}}' = \dot{\boldsymbol{M}}\widehat{\boldsymbol{G}}^{-1}\dot{\boldsymbol{M}}'. \tag{103}$$

Let $\hat{\boldsymbol{d}} = \hat{\boldsymbol{e}} - \boldsymbol{e}_g$, the difference between the empirical Bayes and true conditional expectations $E\{\theta \mid x\}$. The empirical Bayes regret is

$$\sum_{k=1}^{K} f_k(\hat{e}_k - e_{g_k})^2 = \sum_{k=1}^{K} f_k \hat{d}_k^2, \tag{104}$$

as in (40). Under the assumption that $\hat{\boldsymbol{e}}$ is unbiased for $\boldsymbol{e}_g$, the expected EBregret is

$$\sum_{k=1}^{K} f_k \operatorname{Var}\left(\hat{d}_k\right). \tag{105}$$

Tweedie's formula (24) says that

$$\hat{\boldsymbol{d}} = \hat{\boldsymbol{l}}^{\cdot} - \dot{\boldsymbol{l}}, \tag{106}$$

so $\operatorname{cov}(\hat{\boldsymbol{d}}) = \operatorname{cov}(\hat{\boldsymbol{l}}^{\cdot})$. Substituting $\hat{f}_k$ for f_k and $\widehat{\operatorname{cov}}(\hat{\boldsymbol{l}}_k^{\cdot})_{kk}$ for $\operatorname{Var}(\hat{d}_k)$ in (105) gives

$$\widehat{E}\{\text{EBregret}\} = \sum_{k=1}^{K} \hat{f}_k \dot{M}_k' \widehat{\boldsymbol{G}}^{-1} \dot{M}_k, \tag{107}$$

$\dot{M}_k'$ the kth row of $\dot{\boldsymbol{M}}$, and so

$$\begin{aligned} \widehat{E}\{\text{EBregret}\} &= \sum_{k-1}^{K} \hat{f}_k \operatorname{trace}\left(\dot{M}_k' \widehat{\boldsymbol{G}}^{-1} \dot{M}_k\right) = \operatorname{trace}\left(\sum_{k=1}^{K} \hat{f}_k \dot{M}_k' \widehat{\boldsymbol{G}}^{-1} \dot{M}_k\right) \\ &= \operatorname{trace}\left(\widehat{\boldsymbol{G}}^{-1} \sum_{k=1}^{K} \dot{M}_k \hat{f}_k \dot{M}_k'\right) \\ &= \operatorname{trace}\left(\widehat{\boldsymbol{G}}^{-1} \dot{\boldsymbol{M}}' \operatorname{diag}\left(\hat{\boldsymbol{f}}\right) \dot{\boldsymbol{M}}\right), \end{aligned} \tag{108}$$

which is (47).

D. Truncation Suppose x is observed (and known to have occurred) only if $x \in A$, some predetermined subset of the original sample space $\mathcal{X}$, this being the definition of data *truncation*. This changes the marginal density from $f(x)$ to $f_A(x) = f(x)/\pi$ for $x \in A$, with $\pi = \int_A f(y)\,dy$, making a corresponding change in Robbins' estimate (50),

$$e_g(x) = (x+1)f_A(x+1)/f_A(x). \tag{109}$$

Truncation also changes the prior density $g(\theta)$ — see Remark G — accounting for the change in Robbins' rule. Since now $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$ is a random sample from $f_A(x)$, maximum likelihood methods such as (57) will correctly estimate (109).

Truncation affects the distribution $p(x \mid \theta)$ (1),

$$p_A(x \mid \theta) = p(x \mid \theta)/P(\theta) \qquad (P(\theta) = \Pr\{x \in A \mid \theta\}). \tag{110}$$

The truncated version of an exponential family density (22) is

$$p_A(x \mid \theta) = e^{\theta x - (\psi(\theta) + \log P(\theta))} p_0(x), \tag{111}$$

a different exponential family but one having the same base density $p_0(x)$, and therefore the same function $l_0(x) = \log p_0(x)$. The truncated version of Tweedie's formulas (24) are

$$e_g(x) = \dot{l}_A(x) + \dot{l}_0(x) \quad \text{and} \quad v_g(x) = \ddot{l}_A(x) + \ddot{l}_0(x). \tag{112}$$

In the estimated version $\hat{e}_A(x) = \hat{\dot{l}}_A(x) + \dot{l}_0(x)$, the second term is the same as that for the untruncated $\hat{e}_g(x)$, while the first term is estimated directly from $\boldsymbol{x}$, without specific attention to A. The same goes for $\hat{v}_A(x)$. In particular, expression (66) can be used just as stated.

E. g-modeling The details of g-modeling are spelled out in Efron (2016). Here is a brief description pertaining to the simplified situation where both the θ space $\mathcal{T}$ and the x space $\mathcal{X}$ are finite and discrete,

$$\mathcal{T} = \{\theta_{(1)}, \theta_{(2)}, \ldots, \theta_{(m)}\} \quad \text{and} \quad \mathcal{X} = \{x_{(1)}, x_{(2)}, \ldots, x_{(K)}\}. \tag{113}$$

(Continuous x_i's such as those in the gamnormal example of Section 6 are discretized by binning (33).) The g-model consists of a p-parameter exponential family,

$$\boldsymbol{g} = e^{\boldsymbol{Q}\beta - \phi(\beta)}, \tag{114}$$

$\boldsymbol{Q}$ a given $m \times p$ structure matrix having rows q_j', β an unknown p-dimensional parameter vector, and $\phi(\beta) = \log(\sum e^{q_j \beta})$.

For the butterfly analysis, $\mathcal{X} = \{1, 2, \ldots, 24\}$ and $\theta_{(j)} = \exp(\lambda_{(j)})$, with

$$\boldsymbol{\lambda}_{()} = \{-3, -2.8, -2.6, \ldots, 4\}. \tag{115}$$

The gamnormal examples used

$$\mathcal{X} = \{-1.6, -1.4, \ldots, 8.0\} \quad \text{and} \quad \mathcal{T} = \{0, 0.2, 0.4, \ldots, 7.0\}. \tag{116}$$

Both examples used natural spline models with five degrees of freedom, $\boldsymbol{Q} = \text{ns}(\mathcal{T}, \text{ds} = 5)$ (and including a column of ones in $\boldsymbol{Q}$).

Let $\boldsymbol{P}$ be the $K \times m$ matrix (p_{kj}) where

$$p_{kj} = \Pr\{x = x_{(k)} \mid \theta = \theta_{(j)}\}. \tag{117}$$

The marginal density $\boldsymbol{f}(\beta)$ induced by $\boldsymbol{g}(\beta)$ is

$$\boldsymbol{f}(\beta) = \boldsymbol{P}\boldsymbol{g}(\beta). \tag{118}$$

The count vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_K)$, $y_k = \#\{x_i = x_{(k)}\}$, is a sufficient statistic having a K-category multinomial distribution

$$\boldsymbol{y} \sim \text{Mult}_K(N, \boldsymbol{f}(\beta)). \tag{119}$$

Estimation of β from (119) is obtained by penalized maximum likelihood,

$$\hat{\beta} = \arg\max \left\{ \left(\sum_{k=1}^{K} \log f_k(\beta) \right) - c_0 \left(\sum_{1}^{p} \beta_l^2 \right)^{1/2} \right\}, \tag{120}$$

$c_0 = 0.1$ for the butterfly data and 1.0 for the gamnormal examples.

F. g-modeling estimated regret Suppose γ is a function of θ, $\gamma = h(\theta)$, and we are interested in estimating its posterior expectation,

$$E^{(\gamma)}(x) = E\{h(\theta) \mid x\}. \tag{121}$$

Continuing in the discrete setup (113), define $h_j = h(\theta_{(j)})$, $E_k = E^{(\gamma)}(x = x_{(k)})$, etc.,

$$\begin{aligned} &u_{kj} = h_j p_{kj}, && v_{kj} = p_{kj}, \\ \text{and} \quad &\bar{u}_k = \sum_{j=1}^{m} u_{kj} g_k, && \bar{v} = \sum_{j=1}^{m} v_{kj} g_j \end{aligned} \tag{122}$$

($\bar{v}_k = f_k$). Then Bayes rule gives

$$E_k = \bar{u}_k / \bar{v}_k. \tag{123}$$

If $\hat{\boldsymbol{g}}$ is an estimate of $\boldsymbol{g}$, the estimate $\widehat{E}_k$ equals

$$\begin{aligned} \widehat{E}_k &= \frac{\sum_{j=1}^{m} u_{ij} \hat{g}_j}{\sum_{j=1}^{m} v_{ij} \hat{g}_j} = E_k \, \frac{1 + \sum_j \frac{u_{kj}}{\bar{u}_k} (\hat{g}_j - g_j)}{1 + \sum_j \frac{v_{kj}}{\bar{v}_k} (\hat{g}_j - g_j)} \\ &\doteq E_k + \sum_{j=1}^{m} T_{kj} (\hat{g}_j - g_j), \end{aligned} \tag{124}$$

where

$$T_{kj} = E_k \left(\frac{u_{kj}}{\bar{u}_k} - \frac{v_{kj}}{\bar{v}_k} \right). \tag{125}$$

Corollary 1 of Efron (2016) gives delta method approximations for the bias vector and covariance matrix of $\hat{\boldsymbol{g}}$,

$$\hat{\boldsymbol{g}} - \boldsymbol{g} \dot{\sim} (\boldsymbol{B}_g, \boldsymbol{C}_g), \tag{126}$$

based on model (114) and (119). Letting $\boldsymbol{T}$ be the $K \times m$ matrix (T_{kj}) (124) then gives approximate bias and covariance for $\widehat{\boldsymbol{E}}$ as an estimate of $\boldsymbol{E}$,

$$\widehat{\boldsymbol{E}} - \boldsymbol{E} \dot{\sim} (\boldsymbol{T}\boldsymbol{B}_g, \boldsymbol{T}\boldsymbol{C}_g\boldsymbol{T}'). \tag{127}$$

The frequentist expected root mean square error at $x = x_{(k)}$ is

$$\text{rmse}_k = \left[(\boldsymbol{T}\boldsymbol{B}_g)_k^2 + (\boldsymbol{T}\boldsymbol{C}_g\boldsymbol{T}')_{kk} \right]^{1/2}. \tag{128}$$

The last column in Table 5 came from (128), with $h(\theta) = \log\theta$.

If we take $h(\theta) = \theta$ in (121), i.e., $\gamma = \theta$, then E_k equals $e_k = E\{\theta \mid x = x_{(k)}\}$, and likewise $\widehat{E}_k = \hat{e}_k$. From (42) (with $g = \bar{g}$), we get an approximation for the expected empirical Bayes regret from g-modeling,

$$\widehat{\text{EBregret}} = \sum_{k=1}^{K} f_k \, \text{rmse}_k^2. \tag{129}$$

The last columns of Table 1 and Table 2 came from (129).

G. Formula (70) Suppose that there were actually S butterfly species in Malaysia, each with its own Poisson parameter θ_i, but that Corbet only observed those with $x_i \sim \text{Poi}(\theta_i)$ greater than zero (ignoring truncation for $x_i > 24$). If $g^+(\theta)$ is the density function applying to all S species, then truncation gives the density

$$g(\theta) = c g^+(\theta) \cdot (1 - e^{-\theta}), \tag{130}$$

since $\Pr\{x_i > 0 \mid \theta_i\} = 1 - e^{-\theta_i}$.

The expected total number of species Corbet observed is

$$E\{N\} = S \cdot \int_{\mathcal{T}} g^+(\theta)(1 - e^{-\theta}), \tag{131}$$

leading to the estimate $c = S/N$ in (130). Assuming that the capture occurences of each species follow a Poisson process over time with intensity parameter $\theta_i/2$ —so expected number θ_i in two years — gives

$$E\{\text{new}(t)\} = S \cdot \int_{\mathcal{T}} g^+(\theta) e^{-\theta} (1 - e^{-\theta t/2}) \, d\theta, \tag{132}$$

$e^{-\theta}(1 - e^{-\theta t/2})$ being the probability of not being seen in the first two years and then being seen in the next t years. Together, (130)–(132) give formula (70). The frequentist standard error (71) was obtained using a variant of (128).

H. Figure 11 The 3200 gamnormal x_i's were randomly permuted six times. A version of Figure 11 was calculated for each permutation, and these were averaged to give the final Figure 11. This smoothed out irregularities, though all six graphs looked quite similar.

I. The DTI data The observations x_i (86) are definitely not independent, as nearby brain voxels produce correlated observations. Correlation doesn't affect the values of g-modeling or f-modeling estimates, but it does affect their accuracy. In Table 8, the values in rows 1, 2, and 4 remain plausible, but the frequentist standard errors in row 3 are too small.

J. The opt-out analysis It could be argued that taking $f_A(\cdot) = \tilde{f}(\cdot)$ in (95) errs since $\tilde{f}(\cdot)$ assesses the density of *all* the x_i's including those from $f_B(\cdot)$. Using $h_A f_A(x) = f(x) - h_B f_B(x)$, a second iteration of (95) was carried out, this time with

$$\tilde{f}_A(x) = \frac{\tilde{f}(x) - 0.088/\hat{c}}{0.912}. \tag{133}$$

It gave $\hat{h}_B = 0.090$. Subsequent interations made little difference.

REFERENCES

Carlin, B. P. and Gelfand, A. E. (1991). A sample reuse method for accurate parametric empirical Bayes confidence intervals. *J. Roy. Statist. Soc. B* **53** 189–200.

Deely, J. J. and Lindley, D. V. (1981). Bayes empirical Bayes. *J. Amer. Statist. Assoc.* **76** 833–841. MR650894

Efron, B. (1996). Empirical Bayes methods for combining likelihoods. *J. Amer. Statist. Assoc.* **91** 538–565. MR1395725

Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. Institute of Mathematical Statistics Monographs* **1**. Cambridge University Press, Cambridge. MR2724758

Efron, B. (2011). Tweedie's formula and selection bias. *J. Amer. Statist. Assoc.* **106** 1602–1614.

Efron, B. (2014). Two modeling strategies for empirical Bayes estimation. *Statist. Sci.* **29** 285–301.

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

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge University Press, Cambridge. Institute of Mathematical Statistics Monographs (Book 5).

Efron, B. and Morris, C. (1972). Limiting the risk of Bayes and empirical Bayes estimators. II. The empirical Bayes case. *J. Amer. Statist. Assoc.* **67** 130–139. MR0323015

Fisher, R. A., Corbet, A. S. and Williams, C. B. (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. J. and Toulmin, G. H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* **43** 45–63. MR0077039

Gu, J. and Koenker, R. (2016). On a problem of Robbins. *Int. Statist. Rev.* **84** 224–244. MR3537154

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning. Data mining, Inference, and Prediction*, second ed. *Springer Series in Statistics*. Springer, New York. MR2722294

Jiang, W. and Zhang, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist.* **37** 1647–1684. MR2533467

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27** 887–906. MR0086464

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *J. Amer. Statist. Assoc.* **73** 805–811. MR521328

Laird, N. M. and Louis, T. A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *J. Amer. Statist. Assoc.* **82** 739–757. MR909979

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *J. Amer. Statist. Assoc.* **78** 47–65. MR696849

Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950* 131–148. UC Press. MR0044803

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* 157–163. UC Press. MR0084919

Zhang, C.-H. (2003). Compound decision theory and empirical Bayes methods. *Ann. Statist.* **31** 379–390. MR1983534