Statistical Science
2014, Vol. 29, No. 2, 285–301
DOI: 10.1214/13-STS455

Two Modeling Strategies for Empirical Bayes Estimation

Bradley Efron

Abstract. Empirical Bayes methods use the data from parallel experiments, for instance, observations $X_k \sim \mathcal{N}(\Theta_k, 1)$ for $k = 1, 2, \dots, N$, to estimate the conditional distributions $\Theta_k | X_k$. There are two main estimation strategies: modeling on the θ space, called "g-modeling" here, and modeling on the x space, called "f-modeling." The two approaches are described and compared. A series of computational formulas are developed to assess their frequentist accuracy. Several examples, both contrived and genuine, show the strengths and limitations of the two strategies.

Key words and phrases: f-modeling, g-modeling, Bayes rule in terms of f, prior exponential families.

1. INTRODUCTION

Empirical Bayes methods, though of increasing use, still suffer from an uncertain theoretical basis, enjoying neither the safe haven of Bayes theorem nor the steady support of frequentist optimality. Their rationale is often reduced to inserting more or less obvious estimates into familiar Bayesian formulas. This conceals the essential empirical Bayes task: learning an appropriate prior distribution from ongoing statistical experience, rather than knowing it by assumption. Efficient learning requires both Bayesian and frequentist modeling strategies. My plan here is to discuss such strategies in a mathematically simplified framework that, hopefully, renders them more transparent. The development proceeds with some methodological discussion supplemented by numerical examples.

A wide range of empirical Bayes applications have the following structure: repeated sampling from an unknown prior distribution $g(\theta)$ yields unseen realizations

$$\Theta_1, \Theta_2, \dots, \Theta_N. \tag{1.1}$$

Each Θ_k in turn provides an observation $X_k \sim f_{\Theta_k}(\cdot)$ from a known probability family $f_\theta(x)$,

$$X_1, X_2, \dots, X_N. \tag{1.2}$$

On the basis of the observed sample (1.2), the statistician wishes to approximate certain Bayesian inferences that would be directly available if $g(\theta)$ were known. This is the empirical Bayes framework developed and named by Robbins (1956). Both Θ and X are usually one-dimensional variates, as they will be in our examples, though that is of more applied than theoretical necessity.

A central feature of empirical Bayes estimation is that the data arrives on the x scale but inferences are calculated on the θ scale. Two main strategies have developed: modeling on the θ scale, called *g-modeling* here, and modeling on the x scale, called *f-modeling*. G-modeling has predominated in the theoretical empirical Bayes literature, as in Laird (1978), Morris (1983), Zhang (1997), and Jiang and Zhang (2009). Applications, on the other hand, from Robbins (1956) onward, have more often relied on f-modeling, recently as in Efron (2010, 2011) and Brown, Greenshtein and Ritov (2013).

We begin Section 2 with a discretized statement of Bayes theorem that simplifies the nonparametric f-modeling development of Section 3. Parameterized f-modeling, necessary for efficient empirical Bayes estimation, is discussed in Section 4. Section 5 introduces an exponential family class of g-modeling procedures. Classic empirical Bayes applications, an f-modeling stronghold (including Robbins' Poisson formula, the James–Stein estimator and false discovery rate methods), are the subject of Section 6. The paper concludes with a brief discussion in Section 7.

Bradley Efron is Professor of Statistics and Biostatistics, Department of Statistics, Stanford University, Stanford, California 94305-4065, USA (e-mail: brad@stat.stanford.edu).

Several numerical examples, both contrived and genuine, are carried through in Sections 2 through 7. The comparison is never one-sided: as one moves away from the classic applications, g-modeling comes into its own. Trying to go backward, from observations on the x-space to the unknown prior $g(\theta)$, has an ill-posed computational flavor. Empirical Bayes calculations are inherently fraught with difficulties, making both of the modeling strategies useful. An excellent review of empirical Bayes methodology appears in Chapter 3 of Carlin and Louis (2000).

There is an extensive literature, much of it focusing on rates of convergence, concerning the "deconvolution problem," that is, estimating the distribution $g(\theta)$ from the observed X values. A good recent reference is Butucea and Comte (2009). Empirical Bayes inference amounts to estimating certain nonlinear functionals of $g(\cdot)$, whereas linear functionals play a central role for the deconvolution problem, as in Cavalier and Hengartner (2009), but the two literatures are related. The development in this paper employs discrete models that avoid rates of convergence difficulties.

Empirical Bayes analyses often produce impressive-looking estimates of posterior θ distributions. The main results in what follows are a series of computational formulas—Theorems 1 through 4—giving the accuracy of both f-model and g-model estimates. Accuracy can be poor, as some of the examples show, and in any case accuracy assessments are an important part of the analysis.

2. A DISCRETE MODEL OF BAYESIAN INFERENCE

In order to simplify the f-modeling computations, we will assume a model in which both the parameter vector θ and the observed data set x are confined to finite discrete sets:

$$\begin{aligned}&\theta \in \boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_j, \ldots, \theta_m) \quad \text{and}\\ &x \in \mathbf{x} = (x_1, x_2, \ldots, x_i, \ldots, x_n)\end{aligned} \tag{2.1}$$

with $m < n$. The prior distribution $\mathbf{g}$ puts probability g_j on θ_j,

$$\mathbf{g} = (g_1, g_2, \ldots, g_j, \ldots, g_m)'. \tag{2.2}$$

This induces a marginal distribution $\mathbf{f}$ on $\mathbf{x}$,

$$\mathbf{f} = (f_1, f_2, \ldots, f_i, \ldots, f_n)', \tag{2.3}$$

with $f_i = \Pr\{x = x_i\}$. Letting $\{p_{ij}\}$ represent the sampling probabilities

$$p_{ij} = \Pr\{x_i|\theta_j\}, \tag{2.4}$$

the $n \times m$ matrix

$$P = (p_{ij}) \tag{2.5}$$

produces $\mathbf{f}$ from $\mathbf{g}$ according to

$$\mathbf{f} = P\mathbf{g}. \tag{2.6}$$

In the example of Figure 1, we have

$$\boldsymbol{\theta} = (-3, -2.8, \ldots, 3) \quad (m = 31), \tag{2.7}$$

with $g(\theta)$ an equal mixture of a discretized $\mathcal{N}(0, 0.5^2)$ density and a density proportional to $|\theta|$. The sampling probabilities p_{ij} are obtained from the normal translation model $\varphi(x_i - \theta_j)$, φ the standard normal density function, and with

$$\mathbf{x} = (-4.4, -4.35, \ldots, 5.2) \quad (n = 193). \tag{2.8}$$

Then $\mathbf{f} = P\mathbf{g}$ produces the triangular-shaped marginal density $f(x)$ seen in the bottom panel. Looking ahead, we will want to use samples from the bottom distribution to estimate functions of the top.

In the discrete model (2.1)–(2.6), Bayes rule takes the form

$$\Pr\{\theta_j|x_i\} = p_{ij}g_j/f_i. \tag{2.9}$$

Letting $\mathbf{p}_i$ represent the ith row of matrix P, the m-vector of posterior probabilities of θ given $x = x_i$ is given by

$$\operatorname{diag}(\mathbf{p}_i)\mathbf{g}/\mathbf{p}_i\mathbf{g}, \tag{2.10}$$

where diag($\mathbf{v}$) indicates a diagonal matix with diagonal elements taken from the vector $\mathbf{v}$.

Now suppose $t(\theta)$ is a parameter of interest, expressed in our discrete setting by the vector of values

$$\mathbf{t} = (t_1, t_2, \ldots, t_j, \ldots, t_m)'. \tag{2.11}$$

The posterior expectation of $t(\theta)$ given $x = x_i$ is then

$$\begin{aligned}E\{t(\theta)|x_i\} &= \sum_{j=1}^{m} t_j p_{ij} g_j \Big/ f_i\\ &= \mathbf{t}' \operatorname{diag}(\mathbf{p}_i)\mathbf{g}/\mathbf{p}_i\mathbf{g}.\end{aligned} \tag{2.12}$$

The main role of the discrete model (2.1)–(2.6) is to simplify the presentation of f-modeling begun in Section 3. Basically, it allows the use of familiar matrix calculations rather than functional equations. G-modeling, Section 5, will be presented in both discrete and continuous forms. The prostate data example of Section 6 shows our discrete model nicely handling continuous data.

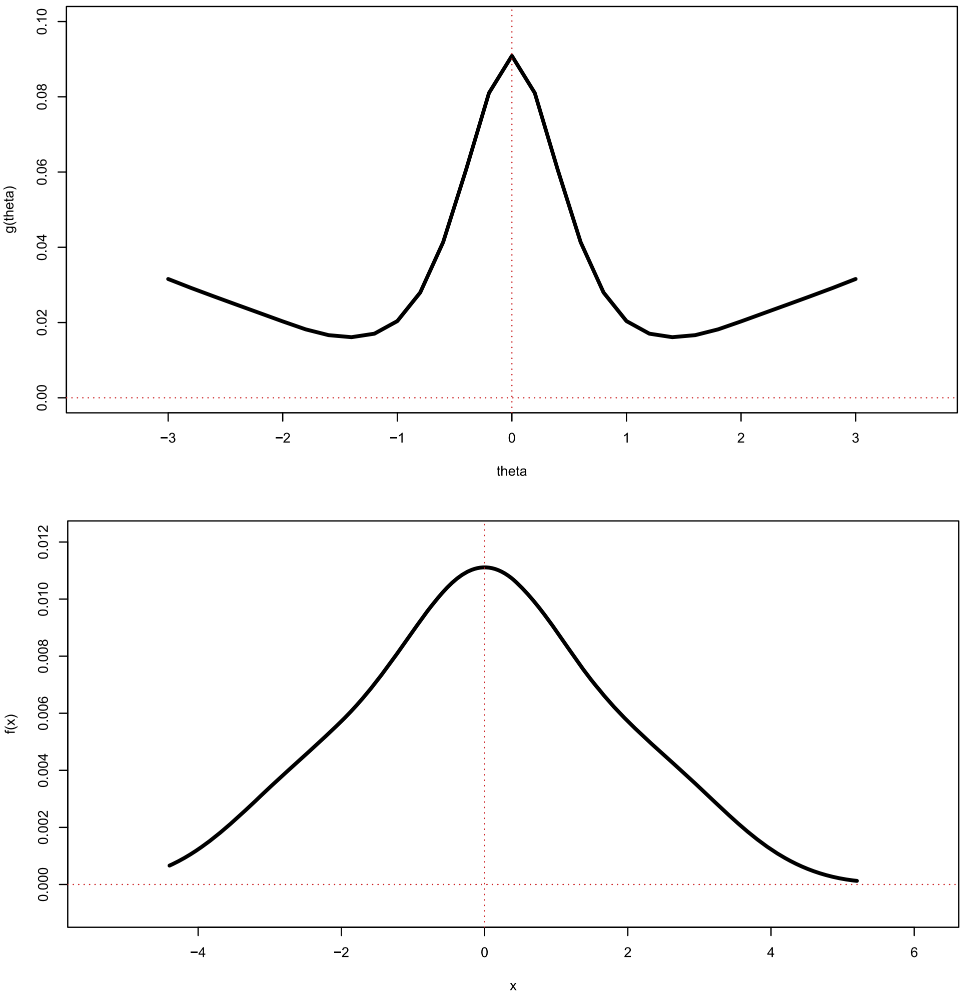

FIG. 1. Top: *Discrete model*: *prior* $g(\theta)$, $\theta = \text{seq}(-3, 3, 0.2)$; *$g$ is equal mixture of* $\mathcal{N}(0, 0.5^2)$ *and density* $\propto |\theta|$. Bottom: *Corresponding* $f(x)$: *assuming* $\mathcal{N}(\theta, 1)$ *sampling*, $x = \text{seq}(-4.4, 5.2, 0.05)$. *Note the different scales.*

3. BAYES RULE IN TERMS OF f

Formula (2.12) expresses $E\{t(\theta)|x_i\}$ in terms of the prior distribution $\mathbf{g}$. This is fine for pure Bayesian applications but in empirical Bayes work, information arrives on the x scale and we may need to express Bayes rule in terms of $\mathbf{f}$. We begin by inverting (2.6), $\mathbf{f} = P\mathbf{g}$.

For now assume that the $n \times m$ matrix P (2.4)–(2.5) is of full rank m. Then the $m \times n$ matrix

$$A = (P'P)^{-1}P' \tag{3.1}$$

carries out the inversion,

$$\mathbf{g} = A\mathbf{f}. \tag{3.2}$$

Section 4 discusses the case where rank(P) is less than m. Other definitions of A are possible; see the discussion in Section 7.

With $\mathbf{p}_i$ denoting the ith row of P as before, let

$$\mathbf{u}' = (\cdots t_j p_{ij} \cdots) = \mathbf{t}' \operatorname{diag}(\mathbf{p}_i), \qquad \mathbf{v}' = \mathbf{p}_i \tag{3.3}$$

and

$$\mathbf{U}' = \mathbf{u}'A, \qquad \mathbf{V}' = \mathbf{v}'A, \tag{3.4}$$

$\mathbf{U}$ and $\mathbf{V}$ being n-vectors. (Here we are suppressing the subscript in $\mathbf{u} = \mathbf{u}_i$, etc.) Using (3.2), the Bayes posterior expectation $E\{t|x_i\}$ (2.12) becomes

$$E\{t|x_i\} = \frac{\mathbf{u}'\mathbf{g}}{\mathbf{v}'\mathbf{g}} = \frac{\mathbf{U}'\mathbf{f}}{\mathbf{V}'\mathbf{f}}, \tag{3.5}$$

the latter being *Bayes rule in terms of* $\mathbf{f}$. Notice that $\mathbf{U}$ and $\mathbf{V}$ do not depend on $\mathbf{g}$ or $\mathbf{f}$. The denominator $\mathbf{V}'\mathbf{f}$ equals $f(x_i)$ in (3.5), but not in the regularized versions of Section 4.

In a typical empirical Bayes situation, as in Section 6.1 of Efron (2010), we might observe independent observations $X_1, X_2, \dots, X_N$ from the marginal density $f(x)$,

$$X_k \overset{\text{i.i.d.}}{\sim} f(\cdot), \quad k = 1, 2, \dots, N, \tag{3.6}$$

and wish to estimate $E = E\{t|x_i\}$. For the discrete model (2.1), the vector of counts $\mathbf{y} = (y_1, y_2, \dots, y_n)'$,

$$y_i = \#\{X_k = x_i\}, \tag{3.7}$$

is a nonparametric sufficient statistic; $\mathbf{y}$ follows a multinomial distribution on n categories, N draws, probability vector $\mathbf{f}$,

$$\mathbf{y} \sim \text{Mult}_n(N, \mathbf{f}), \tag{3.8}$$

having mean vector and covariance matrix

$$\mathbf{y} \sim \big(N\mathbf{f}, ND(\mathbf{f})\big), \quad D(\mathbf{f}) \equiv \text{diag}(\mathbf{f}) - \mathbf{f}\mathbf{f}'. \tag{3.9}$$

The unbiased estimate of $\mathbf{f}$,

$$\hat{\mathbf{f}} = \mathbf{y}/N, \tag{3.10}$$

gives a nonparametric estimate $\hat{E}$ of $E\{t|x_i\}$ by substitution into (3.5),

$$\hat{E} = \mathbf{U}'\hat{\mathbf{f}}/\mathbf{V}'\hat{\mathbf{f}}. \tag{3.11}$$

Using $\hat{\mathbf{f}} \sim (\mathbf{f}, D(\mathbf{f})/N)$, a standard differential argument yields the approximate "delta method" frequentist standard error of $\hat{E}$. Define

$$U_f = \sum_{i=1}^n f_i U_i, \quad V_f = \sum_{i=1}^n f_i V_i \tag{3.12}$$

and

$$\mathbf{W} = \frac{\mathbf{U}}{U_f} - \frac{\mathbf{V}}{V_f}. \tag{3.13}$$

(Notice that $\sum f_i W_i = 0$.)

THEOREM 1. *The delta-method approximate standard deviation of $\hat{E} = \mathbf{U}'\hat{\mathbf{f}}/\mathbf{V}'\hat{\mathbf{f}}$ is*

$$\text{sd}(\hat{E}) = \frac{1}{\sqrt{N}}|E| \cdot \sigma_f(W), \tag{3.14}$$

where $E = \mathbf{U}'\mathbf{f}/\mathbf{V}'\mathbf{f}$ and

$$\sigma_f^2(W) = \sum_{i=1}^n f_i W_i^2. \tag{3.15}$$

The approximate coefficient of variation $\text{sd}(\hat{E})/|E|$ *of $\hat{E}$ is*

$$\text{cv}(\hat{E}) = \sigma_f(W)/\sqrt{N}. \tag{3.16}$$

PROOF. From (3.5) we compute the joint moments of $\mathbf{U}'\hat{\mathbf{f}}$ and $\mathbf{V}'\hat{\mathbf{f}}$,

$$\begin{pmatrix} \mathbf{U}'\hat{\mathbf{f}} \\ \mathbf{V}'\hat{\mathbf{f}} \end{pmatrix} \sim \left(\begin{pmatrix} U_f \\ V_f \end{pmatrix}, \frac{1}{N} \begin{pmatrix} \sigma_f^2(U) & \sigma_f(U,V) \\ \sigma_f(U,V) & \sigma_f^2(V) \end{pmatrix} \right), \tag{3.17}$$

with $\sigma_f^2(U) = \sum f_i(U_i - U_f)^2, \sigma_f(U,V) = \sum f_i(U_i - U_f)(V_i - V_f)$, and $\sigma_f^2(V) = \sum f_i(V_i - V_f)^2$. Then

$$\begin{aligned} \hat{E} = \frac{\mathbf{U}'\hat{\mathbf{f}}}{\mathbf{V}'\hat{\mathbf{f}}} &= E \cdot \frac{1 + \hat{\Delta}_U}{1 + \hat{\Delta}_V} \\ &\doteq E \cdot (1 + \hat{\Delta}_U - \hat{\Delta}_V), \\ &\left[\hat{\Delta}_U = \frac{\mathbf{U}'\hat{\mathbf{f}} - U_f}{U_f}, \hat{\Delta}_V = \frac{\mathbf{V}'\hat{\mathbf{f}} - V_f}{V_f}\right] \end{aligned} \tag{3.18}$$

so $\text{sd}(\hat{E}^2) \doteq E^2 \,\text{var}(\hat{\Delta}_U - \hat{\Delta}_V)$, which, again using (3.9), gives Theorem 1. □

The trouble here, as will be shown, is that $\text{sd}(\hat{E})$ or $\text{cv}(\hat{E})$ may easily become unmanageably large. Empirical Bayes methods require sampling on the x scale, which can be grossly inefficient for estimating functions of θ.

Hypothetically, the X_k's in (3.6) are the observable halves of pairs (Θ, X),

$$(\Theta_k, X_k) \overset{\text{ind}}{\sim} g(\theta) f_\theta(x), \quad k = 1, 2, \dots, N. \tag{3.19}$$

If the Θ_k's *had* been observed, we could estimate $\mathbf{g}$ directly as $\bar{\mathbf{g}} = (\bar{g}_1, \bar{g}_2, \dots, \bar{g}_m)'$,

$$\bar{g}_j = \#\{\Theta_k = \theta_j\}/N, \tag{3.20}$$

leading to the *direct Bayes estimate*

$$\bar{E} = \mathbf{u}'\bar{\mathbf{g}}/\mathbf{v}'\bar{\mathbf{g}}. \tag{3.21}$$

$\bar{E}$ would usually be less variable than $\hat{E}$ (3.11) (and would automatically enforce possible constraints on E such as monotonicity in x_k). A version of Theorem 1 applies here. Now we define

$$\begin{aligned} u_g = \sum_{j=1}^m g_j u_j, \quad v_g = \sum_{j=1}^m g_j v_j \quad \text{and} \\ \mathbf{w} = \mathbf{u}/u_g - \mathbf{v}/v_g. \end{aligned} \tag{3.22}$$

THEOREM 2. *For direct Bayes estimation (3.21), the delta-method approximate standard deviation of $\bar{E}$ is*

$$\text{sd}(\bar{E}) = \frac{1}{\sqrt{N}}|E| \cdot \sigma_g(w), \tag{3.23}$$

TABLE 1
Standard deviation and coefficient of variation of $E\{t(\theta)|x=2.5\}$ (*for* $N=1$); *for the three parameters* (3.26), *with g and f as in Figure* 1; *sdf from Theorem* 1 (3.14); *sdd for direct Bayes estimation, Theorem* 2 (3.23); *sdx from the regularized f-modeling of Section* 4, *Theorem* 3 (4.8)

		$N^{1/2}$ sd			$N^{1/2}$ cv		
$t(\theta)$	$E\{t\|x=2.5\}$	sdf	sdd	sdx	cvf	cvd	cvx
Parameter (1)	2.00	8.74	3.38	2.83	4.4	1.7	1.4
Parameter (2)	4.76	43.4	13.7	10.4	9.1	2.9	2.2
Parameter (3)	0.03	43.9	0.53	1.24	1371	16	39

where

$$\sigma_g^2(w)=\sum_{j=1}^{m} g_j w_j^2; \tag{3.24}$$

$\bar{E}$ *has approximate coefficient of variation*

$$\mathrm{cv}(\bar{E})=\sigma_g(w)/\sqrt{N}. \tag{3.25}$$

The proof of Theorem 2 is the same as that for Theorem 1.

Table 1 concerns the estimation of $E\{t(\theta)|x=2.5\}$ for the situation shown in Figure 1. Three different parameters $t(\theta)$ are considered:

$$\begin{aligned} &(1) && t(\theta)=\theta, \\ &(2) && t(\theta)=\theta^2, \\ &(3) && t(\theta)=\begin{cases}1, & \text{if } \theta\le 0,\\ 0, & \text{if } \theta>0.\end{cases}\end{aligned} \tag{3.26}$$

In the third case, $E\{t(\theta)|x\}=\Pr\{\theta\le 0|x\}$. Cvf is $\sqrt{N}\,\mathrm{cv}(\hat{E})$ (3.16) so cvf$/\sqrt{N}$ is the approximate coefficient of variation of $\hat{E}$, the nonparametric empirical Bayes estimate of $E\{t(\theta)|x=2.5\}$. Cvd is the corresponding quantity (3.25), available only if we could directly observe the Θ_k values in (3.19), while cvx is a regularized version of $\hat{E}$ described in the next section.

Suppose we wish to bound cv$(\hat{E})$ below some prespecified value c_0, perhaps $c_0=0.1$. Then according to (3.16), we need N to equal

$$N=(\mathrm{cv}_1/c_0)^2, \tag{3.27}$$

where cv$_1$ is the numerator $\sigma_f(W)$ of (3.16), for example, cvf in Table 1. For the three parameters (3.26) and for $c_0=0.1$, we would require $N=1936$, 8281 and 187 million, respectively.

The vector **W** for parameter (3) is seen to take on enormous values in Figure 2, resulting in $\sigma_f(W)=1370.7$ for (3.16). The trouble stems from the abrupt discontinuity of t_3 at $\theta=0$, which destabilizes **U** in

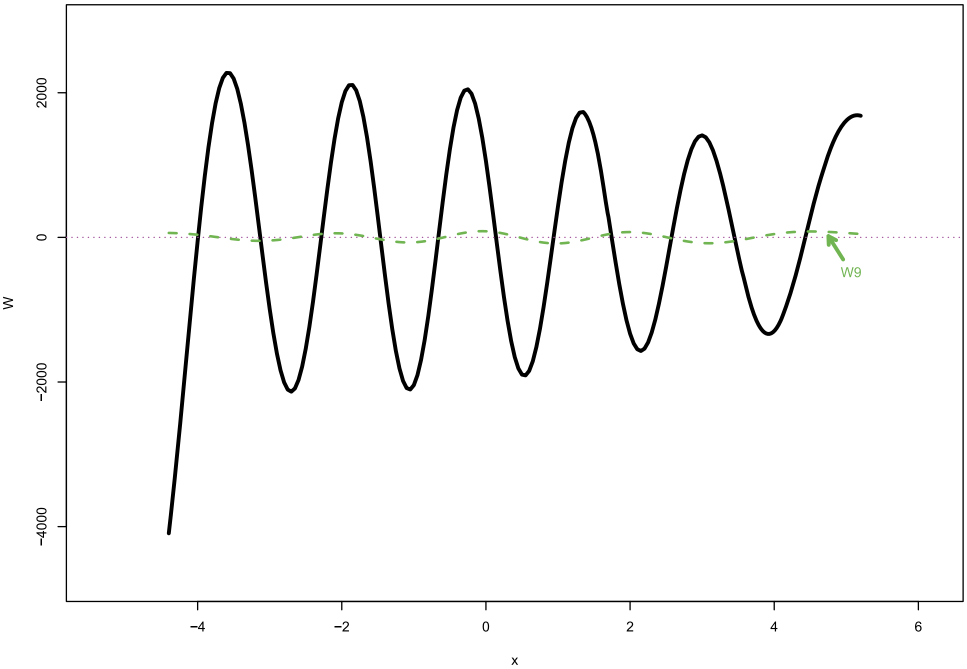

FIG. 2. **W** *vector* (3.13) *for f-Bayes estimation of* $\Pr\{\theta\le 0|x=2.5\}$ *for the model of Figure* 1 (*actually* $\mathbf{W}_{12}$ *as in Section* 4; *dashed curve is* $\mathbf{W}_9$).

(3.13). Definition (3.4) implies $\mathbf{U}'P = \mathbf{u}'$. This says that $\mathbf{U}'$ must linearly compose $\mathbf{u}'$ from the rows of P. But in our example the rows of P are smooth functions of the form $\varphi(x_i - \theta_j)$, forcing the violent cycling of U seen in Figure 2. Section 4 discusses a regularization method that greatly improves the accuracy of using "Bayes rule in terms of **f**."

Table 1 shows that if we *could* sample on the θ scale, as in (3.20), we would require "only" 25,600 Θ_k observations to achieve coefficient of variation 0.1 for estimating $\Pr\{\theta \le 0 | x = 2.5\}$; direct sampling is almost always more efficient than f sampling, but that is not the way empirical Bayes situations present themselves. The efficiency difference is a factor of 86 for parameter (3), but less than a factor of 3 for parameter (1), $t(\theta) = \theta$. The latter is a particularly favorable case for empirical Bayes estimation, as discussed in Section 6.

The assumption of independent sampling, (3.6) and (3.19), is a crucial element of all our results. Independence assumptions (often tacitly made) dominate the empirical Bayes literature, as in Muralidharan et al. (2012), Zhang (1997), Morris (1983), and Efron and Morris (1975). Nonindependence effectively reduces the effective sample size N; see Chapter 8 of Efron (2010). This point is brought up again in Section 6.

4. REGULARIZED f-MODELING

Fully nonparametric estimation of $E = E\{t(\theta)|x\}$ is sometimes feasible, but, as seen in Table 1 of Section 3, it can become unacceptably noisy. Some form of regularization is usually necessary. A promising approach is to estimate **f** parametrically according to a smooth low-dimensional model.

Suppose then that we have such a model, yielding $\hat{\mathbf{f}}$ as an estimate of **f** (2.3), with mean vector and covariance matrix

$$\hat{\mathbf{f}} \sim (\mathbf{f}, \Delta(\mathbf{f})/N). \tag{4.1}$$

In the nonparametric case (3.9) $\Delta(\mathbf{f}) = D(\mathbf{f})$, but we expect that we can reduce $\Delta(\mathbf{f})$ parametrically. In any case, the delta-method approximate coefficient of variation for $\hat{E} = \mathbf{U}'\hat{\mathbf{f}}/\mathbf{V}'\hat{\mathbf{f}}$ (3.11) is given in terms of **W** (3.13):

$$\operatorname{cv}(\hat{E}) = \{\mathbf{W}'\Delta(\mathbf{f})\mathbf{W}/N\}^{1/2}. \tag{4.2}$$

This agrees with (3.16) in the nonparametric situation (3.9) where $\Delta(\mathbf{f}) = \operatorname{diag}(\mathbf{f}) - \mathbf{f}\mathbf{f}'$. The verification of (4.2) is almost identical to that for Theorem 1.

Poisson regression models are convenient for the smooth parametric estimation of **f**. Beginning with an $n \times p$ structure matrix **X**, having rows $\mathbf{x}_i$ for $i = 1, 2, \dots, n$, we assume that the components of the count vector **y** (3.7) are independent Poisson observations,

$$y_i \overset{\text{ind}}{\sim} \operatorname{Poi}(\mu_i), \quad \mu_i = e^{\mathbf{x}_i\alpha} \quad \text{for } i = 1, 2, \dots, n, \tag{4.3}$$

where α is an unknown vector of dimension p. Matrix **X** is assumed to have as its first column a vector of 1's.

Let $\mu_+ = \sum_1^n \mu_i$ and $N = \sum_1^n y_i$, and define

$$f_i = \mu_i/\mu_+ \quad \text{for } i = 1, 2, \dots, n. \tag{4.4}$$

Then a well-known Poisson/multinomial relationship says that the conditional distribution of **y** given N is

$$\mathbf{y}|N \sim \operatorname{Mult}_n(N, \mathbf{f}) \tag{4.5}$$

as in (3.8). Moreover, under mild regularity conditions, the estimate $\hat{\mathbf{f}} = \mathbf{y}/N$ has asymptotic mean vector and covariance matrix (as $\mu_+ \to \infty$)

$$\hat{\mathbf{f}} \dot{\sim} (\mathbf{f}, \Delta(\mathbf{f})/N), \tag{4.6}$$

where

$$\Delta(\mathbf{f}) = \operatorname{diag}(\mathbf{f})\mathbf{X}G_f^{-1}\mathbf{X}'\operatorname{diag}(\mathbf{f}) \quad [G_f = \mathbf{X}'\operatorname{diag}(\mathbf{f})\mathbf{X}]; \tag{4.7}$$

Equations (4.6)–(4.7) are derived from standard generalized linear model calculations. Combining (4.2) and (4.6) gives a Poisson regression version of Theorem 1.

THEOREM 3. *The delta-method coefficient of variation for $\hat{E} = \mathbf{U}'\hat{\mathbf{f}}/\mathbf{V}'\hat{\mathbf{f}}$ under Poisson model* (4.3) *is*

$$\operatorname{cv}(\hat{E}) = \{(\mathbf{W}'\mathbf{X})_f(\mathbf{X}'\mathbf{X})_f^{-1}(\mathbf{W}'\mathbf{X})_f'/N\}^{1/2}, \tag{4.8}$$

where

$$(\mathbf{W}'\mathbf{X})_f = \mathbf{W}'\operatorname{diag}(\mathbf{f})\mathbf{X} \quad \textit{and} \quad (\mathbf{X}'\mathbf{X})_f = \mathbf{X}'\operatorname{diag}(\mathbf{f})\mathbf{X}, \tag{4.9}$$

with **W** *as in* (3.13).

The bracketed term in (4.8), times N, is recognized as the length^2 of the projection of **W** into the p-dimensional space spanned by the columns of **X**, carried out using inner product $\langle a, b\rangle_f = \sum f_i a_i b_i$. In the nonparametric case, **X** equals the identity I, and (4.8) reduces to (3.16). As in (3.14), $\operatorname{sd}(\hat{E})$ is approximated by $|E|\operatorname{cv}(\hat{E})$. [*Note*: Theorem 3 remains valid as stated if a multinomial model for $\hat{\mathbf{f}}$ replaces the Poisson calculations in (4.7).]

Cvx in Table 1 was calculated as in (4.8), with $N = 1$. The structure matrix $\mathbf{X}$ for the example in Figure 1 was obtained from the R natural spline function $ns(x, df = 5)$; including a column of 1's made $\mathbf{X}193 \times 6$. The improvements over cvf, the nonparametric coefficients of variation, were by factors of 3, 5 and 100 for the three parameters (3.26).

The regularization in Theorem 3 takes place with respect to $\mathbf{f}$ and $\hat{\mathbf{f}}$. Good performance also requires regularization of the inversion process $\hat{\mathbf{g}} = A\hat{\mathbf{f}}$ (3.2). Going back to the beginning of Section 3, let

$$P = LDR' \tag{4.10}$$

represent the singular value decomposition of the $n \times m$ matrix P, with L the $n \times m$ orthonormal matrix of left singular vectors, R the $m \times m$ orthonormal matrix of right singular vectors, and D the $m \times m$ diagonal matrix of singular values,

$$d_1 \geq d_2 \geq \cdots \geq d_m. \tag{4.11}$$

Then it is easy to show that the $m \times n$ matrix

$$A = RD^{-1}L' \tag{4.12}$$

is the *pseudo-inverse* of P, which is why we could go from $\mathbf{f} = P\mathbf{g}$ to $\mathbf{g} = A\mathbf{f}$ at (3.2). [Other pseudo-inverses exist; see (7.1).]

Definition (4.12) depends on P being of full rank m, equivalently having $d_m > 0$ in (4.11). Whether or not this is true, very small values of d_j will destabilize A. The familiar cure is to truncate representation (4.12), lopping off the end terms of the singular value decomposition. If we wish to stop after the first r terms, we define R_r to be the first r columns of R, L_r the first r columns of L, D_r the $r \times r$ diagonal matrix $\text{diag}(d_1, d_2, \ldots, d_r)$, and

$$A_r = R_r D_r^{-1} L_r'. \tag{4.13}$$

In fact, $r = 12$ was used in Figure 2 and Table 1, chosen to make

$$\sum_{r+1}^{m} d_j^2 \Big/ \sum_{1}^{m} d_j^2 < 10^{-10}. \tag{4.14}$$

As in (3.1)–(3.13), let

$$\mathbf{U}_r' = \mathbf{u}' A_r, \quad \mathbf{V}_r' = \mathbf{v}' A_r \tag{4.15}$$

[$\mathbf{u}$ and $\mathbf{v}$ stay the same as in (3.3)],

$$E_r = \frac{\mathbf{U}_r'\mathbf{f}}{\mathbf{V}_r'\mathbf{f}}, \quad \hat{E}_r = \frac{\mathbf{U}_r'\hat{\mathbf{f}}}{\mathbf{V}_r'\hat{\mathbf{f}}} \tag{4.16}$$

TABLE 2

Coefficient of variation and standard deviation ($N = 1$), *for* $E\{t|x = 2.5\}$ *as in* 1; *now using Poisson regression in Theorem* 3, *with* $\mathbf{X}$ *based on a natural spline with* 5 *degrees of freedom. Increasing choice of* r, (4.13)–(4.17), *decreases bias but increases variability of* $\hat{E}$ *for parameter* (3); *g error from* (4.20)

		Parameter (1)			**Parameter (3)**		
r	**g error**	**E_r**	**cvx**	**sdx**	**E_r**	**cvx**	**sdx**
3	0.464	1.75	1.00	1.75	0.021	3.6	0.1
6	0.254	2.00	1.34	2.68	0.027	4.6	0.1
9	0.110	2.00	1.36	2.73	0.031	8.2	0.3
12	0.067	2.00	1.41	2.83	0.032	38.6	1.2
15	0.024	2.00	1.39	2.78	0.033	494.0	16.1
18	0.012	2.00	1.39	2.78	0.033	23,820.8	783.8
21	0.006	2.00	1.40	2.80	0.033	960,036.4	31,688.8

and

$$\mathbf{W}_r = \frac{\mathbf{U}_r}{\sum f_i U_{ri}} - \frac{\mathbf{V}_r}{\sum f_i V_{ri}}. \tag{4.17}$$

Theorem 3 then remains valid, with $\mathbf{W}_r$ replacing $\mathbf{W}$. *Note*: Another regularization method, which will not be pursued here, is the use of ridge regression rather than truncation in the inversion process (3.2), as in Hall and Meister (2007).

Reducing r reduces $\mathbf{W}_r$, hence reducing (4.9) and the approximate coefficient of variation of $\hat{E}_r$. The reduction can be dramatic. W_9 almost disappears compared to W_{12} in Figure 2. Table 2 compares various choices of r for parameters (1) and (3) (3.26). The choice turns out to be unimportant for parameter (1) and crucial for parameter (3).

Why not always choose a small value of r? The trouble lies in possible bias for the estimation of $E = E\{t|x\}$. Rather than the crucial inverse mapping $\mathbf{g} = A\mathbf{f}$ (3.2), we get an approximation

$$\begin{aligned} \mathbf{g}_r &= A_r\mathbf{f} = A_r P\mathbf{g} \\ &= R_r D_r^{-1} L_r' LDR'\mathbf{g} = R_r R_r'\mathbf{g} \end{aligned} \tag{4.18}$$

[the last step following from $LDR' = L_r D_r R_r' + L_{(r)} D_{(r)} R_{(r)}'$, with $L_{(r)}$ indicating the last $m - r$ columns of L, etc.; Equation (4.18) says that $\mathbf{g}_r$ is the projection of $\mathbf{g}$ into the linear space spanned by the first r columns of R]. Then, looking at (4.15)–(4.16),

$$E_r = \frac{\mathbf{U}_r'\mathbf{f}}{\mathbf{V}_r'\mathbf{f}} = \frac{\mathbf{u}'\mathbf{g}_r}{\mathbf{v}'\mathbf{g}_r}, \tag{4.19}$$

possibly making $\hat{E}_r$ badly biased for estimating $E = \mathbf{u}'\mathbf{g}/\mathbf{v}'\mathbf{g}$.

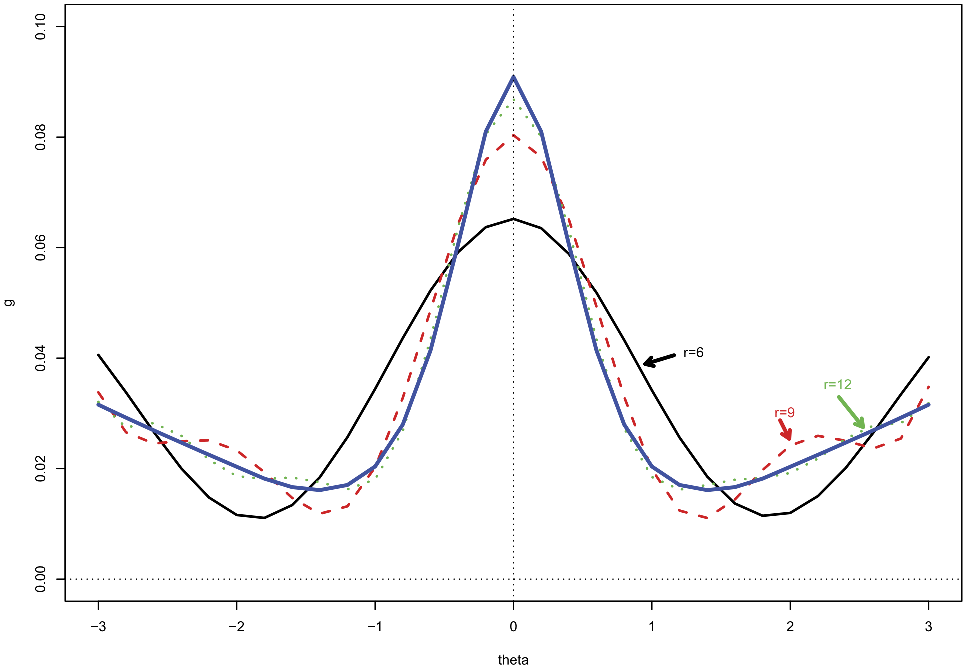

FIG. 3. *Approximation g_r (4.18) with $r = 6, 9, 12$ for g in Figure 1; heavy blue curve is g.*

The E_r columns of Table 2 show that bias is a problem only for quite small values of r. However, the example of Figure 1 is "easy" in the sense that the true prior $\mathbf{g}$ is smooth, which allows $\mathbf{g}_r$ to rapidly approach $\mathbf{g}$ as r increases, as pictured in Figure 3. The g_{error} column of Table 2 shows this numerically in terms of the absolute error

$$g_{\text{error}} = \sum_{i=1}^{m} |g_{ri} - g_i|. \tag{4.20}$$

A more difficult case is illustrated in Figure 4. Here $\mathbf{g}$ is a mixture: 90% of a delta function at $\theta = 0$ and 10% of a uniform distribution over the 31 points θ_j in $\boldsymbol{\theta} = (-3, -2.8, \dots, 3)$; P and $\mathbf{x}$ are as before. Now g_{error} exceeds 1.75 even for $r = 21$; $\mathbf{g}_r$ puts too small a weight on $\theta = 0$, while bouncing around erratically for $\theta \neq 0$, often going negative.

We expect, correctly, that empirical Bayes estimation of $E\{t(\theta)|x\}$ will usually be difficult for the situation of Figure 4. This is worrisome since its $\mathbf{g}$ is a reasonable model for familiar false discovery rate analyses, but see Section 6. Section 5 discusses a different regularization approach that ameliorates, without curing, the difficulties seen here.

5. MODELING THE PRIOR DISTRIBUTION g

The regularization methods of Section 4 involved modeling $\mathbf{f}$, the marginal distribution (2.3) on the x-space, for example, by Poisson regression in Table 2. Here we discuss an alternative strategy: modeling $\mathbf{g}$, the prior distribution (2.2) on the θ-space. This has both advantages and disadvantages, as will be discussed.

We begin with an $m \times q$ model matrix Q, jth row Q_j, which determines $\mathbf{g}$ according to

$$\mathbf{g}(\alpha) = e^{Q\alpha - \mathbf{1}_m \phi(\alpha)} \quad \left[\phi(\alpha) = \log \sum_{1}^{m} e^{Q_j \alpha}\right]. \tag{5.1}$$

[For $\mathbf{v} = (v_1, v_2, \dots, v_m)$, $e^{\mathbf{v}}$ denotes a vector with components e^{v_j}; $\mathbf{1}_m$ is a vector of m 1's, indicating in (5.1) that $\phi(\alpha)$ is subtracted from each component of $Q\alpha$.] Here α is the unknown q-dimensional natural parameter of exponential family (5.1), which determines the prior distribution $\mathbf{g} = \mathbf{g}(\alpha)$. In an empirical Bayes framework, $\mathbf{g}$ gives $\mathbf{f} = P\mathbf{g}$ (2.6), and the statistician then observes a multinomial sample $\mathbf{y}$ of size N from $\mathbf{f}$ as in (3.8),

$$\mathbf{y} \sim \text{Mult}_n(N, P\mathbf{g}(\alpha)), \tag{5.2}$$

from which inferences about $\mathbf{g}$ are to be drawn.

Model (5.1)–(5.2) is not an exponential family in $\mathbf{y}$, a theoretical disadvantage compared to the Poisson modeling of Theorem 3. [It is a *curved exponential family*, Efron (1975).] We can still pursue an asymptotic analysis of its frequentist accuracy. Let

$$D(\mathbf{g}) \equiv \text{diag}(\mathbf{g}) - \mathbf{g}\mathbf{g}', \tag{5.3}$$

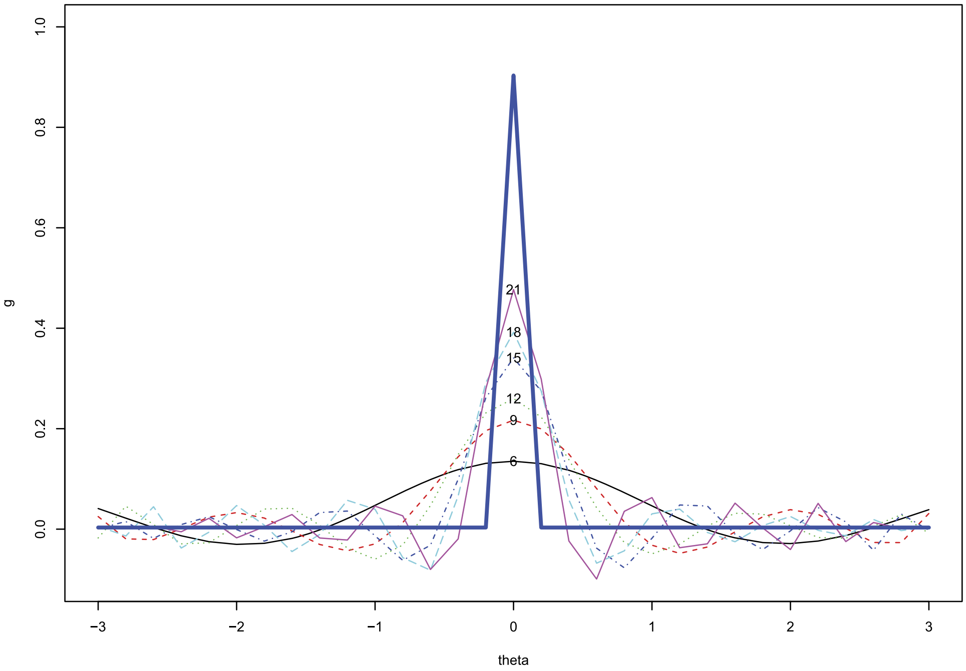

FIG. 4. *True* $g = 0.90 \cdot \delta(0) + 0.10$ *uniform* (*heavy curve*); *approximation* g_r (4.18) *for* $r = 6, 9, 12, 15, 18, 21$, *as labeled.*

the covariance matrix of a single random draw Θ from distribution $\mathbf{g}$, and define

$$Q_\alpha = D(\mathbf{g}(\alpha))Q. \tag{5.4}$$

LEMMA 1. *The Fisher information matrix for estimating* α *in model* (5.1)–(5.2) *is*

$$\mathcal{I} = NQ'_\alpha P' \operatorname{diag}(1/\mathbf{f}(\alpha))PQ_\alpha, \tag{5.5}$$

where P *is the sampling density matrix* (2.5), *and* $\mathbf{f}(\alpha) = P\mathbf{g}(\alpha)$.

PROOF. Differentiating $\log \mathbf{g}$ in (5.1) gives the $m \times q$ derivative matrix $d\log g_i/d\alpha_k$,

$$\frac{d\log \mathbf{g}}{d\alpha} = [I - \mathbf{1}_m \mathbf{g}(\alpha)']Q, \tag{5.6}$$

so

$$\begin{aligned}\frac{d\mathbf{g}}{d\alpha} &= \operatorname{diag}(\mathbf{g}(\alpha))\frac{d\log \mathbf{g}}{d\alpha}\\ &= D(\mathbf{g}(\alpha))Q = Q_\alpha.\end{aligned} \tag{5.7}$$

This yields $d\mathbf{f}/d\alpha = PQ_\alpha$ and

$$\frac{d\log \mathbf{f}}{d\alpha} = \operatorname{diag}\left(\frac{1}{\mathbf{f}(\alpha)}\right)PQ_\alpha. \tag{5.8}$$

The log likelihood from multinomial sample (5.2) is

$$l_\alpha(\mathbf{y}) = \mathbf{y}' \log \mathbf{f}(\alpha) + \text{constant}, \tag{5.9}$$

giving score vector

$$\frac{dl_\alpha(\mathbf{y})}{d\alpha} = \mathbf{y}'\frac{d\log \mathbf{f}}{d\alpha}. \tag{5.10}$$

Since $\mathbf{y}$ has covariance matrix $N(\operatorname{diag}\mathbf{f} - \mathbf{ff}')$ (3.9), $\mathcal{I}$, the covariance matrix of the score vector, equals

$$\begin{aligned}\mathcal{I} &= NQ'_\alpha P' \operatorname{diag}(1/\mathbf{f})(\operatorname{diag}\mathbf{f} - \mathbf{ff}')\\ &\quad \cdot \operatorname{diag}(1/\mathbf{f})PQ_\alpha\\ &= NQ'_\alpha P'(\operatorname{diag}(1/\mathbf{f}) - \mathbf{1}_n\mathbf{1}'_n)PQ_\alpha.\end{aligned} \tag{5.11}$$

Finally,

$$\mathbf{1}'_n PQ_\alpha = \mathbf{1}'_m D(g(\alpha))Q = \mathbf{0}'Q = 0 \tag{5.12}$$

(using the fact that the columns of P sum to 1), and (5.11) yields the lemma. □

Standard sampling theory says that the maximum likelihood estimate (MLE) $\hat{\alpha}$ has approximate covariance matrix $\mathcal{I}^{-1}$ and that $\hat{\mathbf{g}} = \mathbf{g}(\hat{\alpha})$ has approximate covariance, from (5.7),

$$\operatorname{cov}(\hat{\mathbf{g}}) = Q_\alpha \mathcal{I}^{-1} Q'_\alpha. \tag{5.13}$$

LEMMA 2. *The approximate covariance matrix for the maximum likelihood estimate* $\mathbf{g}(\hat{\alpha})$ *of* $\mathbf{g}$ *in model* (5.1)–(5.2) *is*

$$\begin{aligned}&\operatorname{cov}(\hat{\mathbf{g}})\\ &\quad = \frac{1}{N}Q_\alpha[Q'_\alpha P' \operatorname{diag}(1/\mathbf{f}(\alpha))PQ_\alpha]^{-1}Q'_\alpha.\end{aligned} \tag{5.14}$$

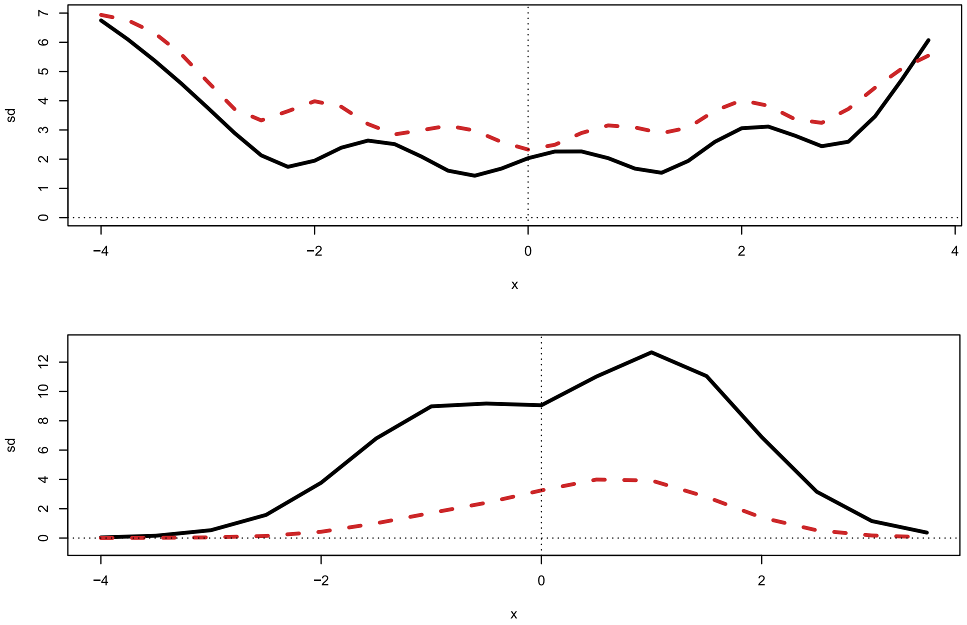

FIG. 5. Top: *Standard deviation of* $E\{t|x\}$ *as a function of* x, *for parameter* (1) $t(\theta)=\theta$ (*with* $N=1$); f*-modeling* (*solid*), g*-modeling* (*dashed*). Bottom: *Now for parameter* (3), $t(\theta)=1$ *or* 0 *as* $\theta\le 0$ *or* >0; *using natural spline models*, $df=6$, *for both calculations.*

If we are interested in a real-valued parameter $\tau = T(\mathbf{g})$, the approximate standard deviation of its MLE $\hat{\tau}=T(g(\hat{\alpha}))$ is

$$\mathrm{sd}(\hat{\tau})=\left[\dot{T}'\operatorname{cov}(\hat{\mathbf{g}})\dot{T}\right]^{1/2}, \tag{5.15}$$

where $\dot{T}$ is the gradient vector $dT/d\mathbf{g}$, evaluated at $\hat{\mathbf{g}}$. When $T(\mathbf{g})$ is the conditional expectation of a parameter $t(\theta)$ (3.5),

$$T(\mathbf{g})=E\{t(\theta)|x=x_i\}=\mathbf{u}'\mathbf{g}/\mathbf{v}'\mathbf{g}, \tag{5.16}$$

we compute

$$\dot{T}(\mathbf{g})=\mathbf{w}=(\mathbf{u}/u_g)-(\mathbf{v}/v_g) \tag{5.17}$$

(3.23), and get the following.

THEOREM 4. *Under models* (5.1)–(5.2), *the MLE* $\hat{E}$ *of* $E\{t(\theta)|x=x_i\}$ *has approximate standard deviation*

$$\mathrm{sd}(\hat{E})=|E|\left[\mathbf{w}'\operatorname{cov}(\hat{\mathbf{g}})\mathbf{w}\right]^{1/2}, \tag{5.18}$$

with $\mathbf{w}$ *as in* (5.17) *and* $\operatorname{cov}(\hat{\mathbf{g}})$ *from* (5.14).

We can now compare $\mathrm{sd}(\hat{E})$ from $\mathbf{g}$-modeling (5.18), with the corresponding $\mathbf{f}$-modeling results of Theorem 3. Figure 5 does this with parameters (1) and (3) (3.26) for the example of Figure 1. Theorem 3, modified as at (4.17) with $r=12$, represents $\mathbf{f}$-modeling, now with X based on $ns(\mathbf{x},6)$, a natural spline with six degrees of freedom. Similarly for $\mathbf{g}$-modeling, $Q=ns(\boldsymbol{\theta},6)$ in (5.1); α was chosen to make $\mathbf{g}(\alpha)$ very close to the upper curve in Figure 1. (Doing so required six rather than five degrees of freedom.)

The upper panel of Figure 5 shows $\mathbf{f}$-modeling yielding somewhat smaller standard deviations for parameter (1), $t(\theta)=\theta$. This is an especially favorable case for $\mathbf{f}$-modeling, as discussed in Section 6. However, for parameter (3), $E=\Pr\{t\le 0|x\}$, $\mathbf{g}$-modeling is far superior. *Note:* in exponential families, curved or not, it can be argued that the effective degrees of freedom of a model equals its number of free parameters; see Remark D of Efron (2004). The models used in Figure 5 each have six parameters, so in this sense the comparison is fair.

Parametric g-space modeling, as in (5.1), has several advantages over the f-space modeling of Section 4:

Constraints. $\hat{\mathbf{g}}=\exp(Q\hat{\alpha}-\mathbf{1}_m\phi(\hat{\alpha}))$ has all coordinates positive, unlike the estimates seen in Figure 4. Other constraints such as monotonicity or convexity that may be imposed on $\hat{\mathbf{f}}=P\hat{\mathbf{g}}$ by the structure of P are automatically enforced, as discussed in Chapter 3 of Carlin and Louis (2000).

Accuracy. With some important exceptions, discussed in Section 6, g-modeling often yields smaller values of $\mathrm{sd}(\hat{E})$, as typified in the bottom panel of Figure 5. This is particularly true for discontinuous parameters $t(\theta)$, such as parameter (3) in Table 1.

Simplicity. The bias/variance trade-offs involved with the choice of r in Section 4 are avoided and, in fact, there is no need for "Bayes rule in terms of $\mathbf{f}$."

Continuous formulation. It is straightforward to translate g-modeling from the discrete framework (2.1)–(2.4) into more familiar continuous language. Exponential family model (5.1) now becomes

$$g_\alpha(\theta) = e^{\mathbf{q}(\theta)\alpha - \phi(\alpha)} \qquad \left[\phi(\alpha) = \log \int e^{\mathbf{q}(\theta)\alpha}\, d\theta\right], \tag{5.19}$$

where $\mathbf{q}(\theta)$ is a smoothly defined $1 \times q$ vector function of θ. Letting $f_\theta(x)$ denote the sampling density of x given θ, define

$$h(x) = \int f_\theta(x) g(\theta)(\mathbf{q}(\theta) - \bar{\mathbf{q}})\, d\theta \qquad \left[\bar{\mathbf{q}} = \int g(\theta)\mathbf{q}(\theta)\, d\theta\right]. \tag{5.20}$$

Then the $q \times q$ information matrix $\mathcal{I}$ (5.5) is

$$\mathcal{I} = N \int \left[\frac{h(x)'h(x)}{f(x)^2}\right] f(x)\, dx \qquad \left[f(x) = \int g(\theta) f_\theta(x)\, dx\right]. \tag{5.21}$$

A posterior expectation $E = E\{t(\theta)|x\}$ has MLE

$$\hat{E} = \int t(\theta) f_\theta(x) g_{\hat{\alpha}}(\theta)\, d\theta \Big/ \int f_\theta(x) g_{\hat{\alpha}}(\theta)\, d\theta. \tag{5.22}$$

An influence function argument shows that E has gradient

$$\frac{dE}{d\alpha} = E \int z(\theta) g_\alpha(\theta)(\mathbf{q}(\theta) - \bar{\mathbf{q}})\, d\theta, \tag{5.23}$$

with

$$z(\theta) = \frac{t(\theta) f_\theta(x) g_\alpha(\theta)}{\int t(\varphi) f_\varphi(x) g_\alpha(\varphi)\, d\varphi} - \frac{f_\theta(x) g_\alpha(\theta)}{\int f_\varphi(x) g_\alpha(\varphi)\, d\varphi}. \tag{5.24}$$

Then the approximate standard deviation of $\hat{E}$ is

$$\mathrm{sd}(\hat{E}) = \left(\frac{dE}{d\alpha} \mathcal{I}^{-1} \frac{dE'}{d\alpha}\right)^{1/2}, \tag{5.25}$$

combining (5.21)–(5.24). [Of course, the integrals required in (5.25) would usually be done numerically, implicitly returning us to discrete calculations!]

Modeling the prior. Modeling on the g-scale is convenient for situations where the statistician has qualitative knowledge concerning the shape of the prior $\mathbf{g}$. As a familiar example, large-scale testing problems often have a big atom of prior probability at $\theta = 0$, corresponding to the null cases. We can accommodate this by including in model matrix Q (5.1) a column $\mathbf{e}_0 = (0, 0, \dots, 0, 1, 0, \dots, 0)'$, with the 1 at $\theta = 0$.

Such an analysis was carried out for the situation in Figure 4, where the true $\mathbf{g}$ equaled $0.9\mathbf{e}_0 + 0.1 \cdot$ uniform. Q was taken to be the natural spline basis $ns(\boldsymbol{\theta}, 5)$ augmented by column $\mathbf{e}_0$, a 31×6 matrix. Table 3 shows the results for $\mathbf{t} = \mathbf{e}_0$, that is, for

$$E = E\{t|x\} = \Pr\{\theta = 0|x\}. \tag{5.26}$$

The table gives E and $\mathrm{sd}(\hat{E})$ (5.18) for $x = -4, -3, \dots, 4$ ($N = 1$), as well as the coefficient of variation $\mathrm{sd}(\hat{E})/E$.

The results are not particularly encouraging: we would need sample sizes N on the order of 10,000 to expect reasonably accurate estimates $\hat{E}$ (3.27). On the other hand, f-modeling as in Section 4 is hopeless here. Section 6 has more to say about false discovery rate estimates (5.26).

A random sample of $N = 5000$ X values was drawn from the distribution $\mathbf{f} = P\mathbf{g}$ corresponding to the true $\mathbf{g}$ in Figure 4 [with P based on the normal density $\varphi(x_i - \theta_j)$ as before], giving count vector $\mathbf{y}$ (3.7). Numerical maximization yielded $\hat{\alpha}$, the MLE in model (5.1)–(5.2), Q as in Table 3. The estimate $\hat{\mathbf{g}} = \mathbf{g}(\hat{\alpha})$ put probability 0.920 at $\theta = 0$, compared to true value 0.903, with nonnull distribution as shown in Figure 6. The nonnull peaks at $\theta = \pm 2$ were artifacts of the estimation procedure. On the other hand, $\hat{\mathbf{g}}$ correctly put

TABLE 3

Estimating $E = \Pr\{\theta = 0|x\}$ in the situation of Figure 4; using g-modeling (5.1) with Q equal $ns(x, 5)$ augmented with a column putting a delta function at $\theta = 0$. Sd is $\mathrm{sd}(\hat{E})$ (5.25), cv is the coefficient of variation sd/E. (For sample size N, divide entries by $N^{1/2}$.)

x	−4	−3	−2	−1	0	1	2	3	4
E	0.04	0.32	0.78	0.94	0.96	0.94	0.78	0.32	0.04
$N^{1/2}\cdot$ sd	0.95	3.28	9.77	10.64	9.70	10.48	9.92	3.36	0.75
$N^{1/2}\cdot$ cv	24.23	10.39	12.53	11.38	10.09	11.20	12.72	10.65	19.21

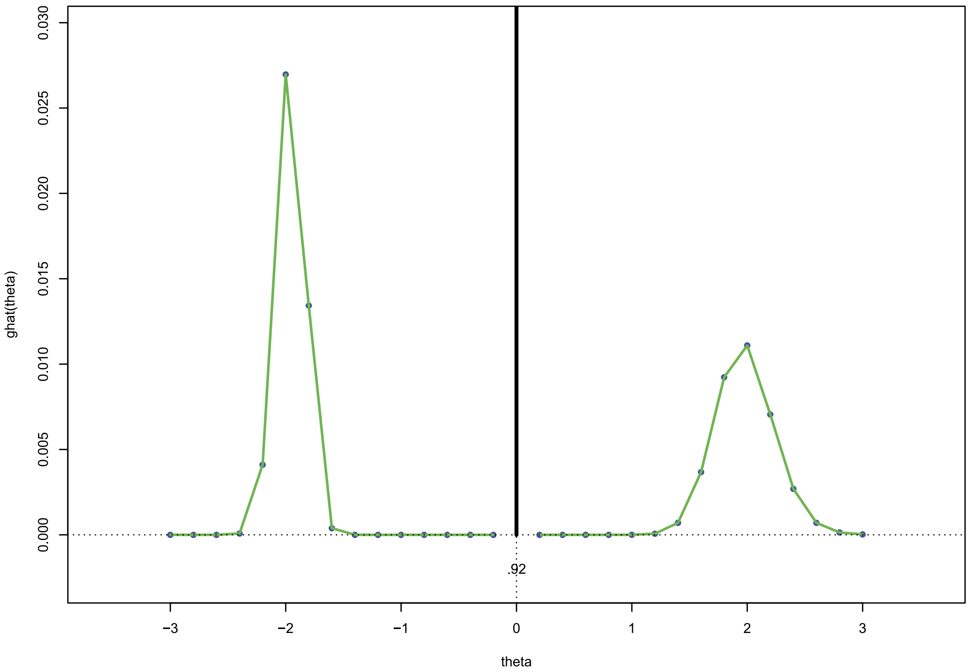

FIG. 6. *MLE nonnull distribution, estimated from a sample of* $N = 5000$ X *values from* f *corresponding to true* g *in Figure* 4; *estimated atom at* $\theta = 0$ *was* 0.92.

roughly equal nonnull probability above and below 0. This degree of useful but crude inference should be kept in mind for the genuine data examples of Section 6, where the truth is unknown.

Our list of g-modeling advantages raises the question of why f-modeling has dominated empirical Bayes applications. The answer—that a certain class of important problems is more naturally considered in the f domain—is discussed in the next section. Theoretically, as opposed to practically, g-modeling has played a central role in the empirical Bayes literature. Much of that work involves the nonparametric maximum likelihood estimation of the prior distribution $g(\theta)$, some notable references being Laird (1978), Zhang (1997) and Jiang and Zhang (2009). Parametric g-modeling, as discussed in Morris (1983) and Casella (1985), has been less well developed. A large part of the effort has focused on the "normal-normal" situation, normal priors with normal sampling errors, as in Efron and Morris (1975), and other conjugate situations. Chapter 3 of Carlin and Louis (2000) gives a nice discussion of parametric empirical Bayes methods, including binomial and Poisson examples.

6. CLASSIC EMPIRICAL BAYES APPLICATIONS

Since its post-war emergence (Robbins, 1956, Good and Toulmin, 1956, James and Stein, 1961), empirical Bayes methodology has focused on a small set of specially structured situations: ones where certain Bayesian inferences can be computed simply and directly from the marginal distribution of the observations on the x-space. There is no need for g-modeling in this framework or, for that matter, any calculation of $\hat{\mathbf{g}}$ at all. False discovery rates and the James–Stein estimator fall into this category, along with related methods discussed in what follows. Though g-modeling is unnecessary here, it will still be interesting to see how it performs on the classic problems.

Robbins' Poisson estimation example exemplifies the classic empirical Bayes approach: independent but not identically distributed Poisson variates

$$X_k \overset{\text{ind}}{\sim} \text{Poi}(\Theta_k), \quad k = 1, 2, \dots, N, \tag{6.1}$$

are observed, with the Θ_k's notionally drawn from some prior $g(\theta)$. Applying Bayes rule with the Poisson kernel $e^{-\theta}\theta^x/x!$ shows that

$$E\{\theta|x\} = (x+1) f_{x+1}/f_x, \tag{6.2}$$

where $\mathbf{f} = (f_1, f_2, \dots)$ is the marginal distribution of the X's. [This is an example of (3.5), Bayes rule in terms of $\mathbf{f}$; defining $\mathbf{e}_i = (0, 0, \dots, 1, 0, \dots, 0)'$ with 1 in the ith place, $\mathbf{U} = (x+1)\mathbf{e}_{x+1}$, and $\mathbf{V} = \mathbf{e}_x$.] Letting $\hat{\mathbf{f}} = (\hat{f}_1, \hat{f}_2, \dots)$ be the nonparametric MLE (3.10), Robbins' estimate is the "plug-in" choice

$$\hat{E}\{\theta|x\} = (x+1) \hat{f}_{x+1}/\hat{f}_x \tag{6.3}$$

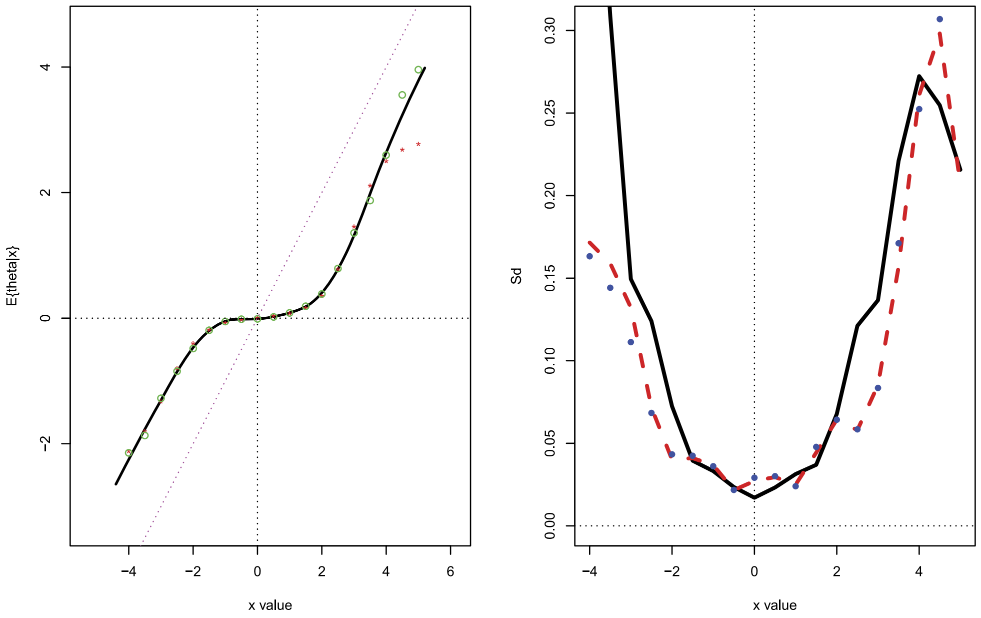

FIG. 7. *Prostate data. Left panel shows estimates of* $E\{\theta|x\}$ *from Tweedie's formula* (*solid curve*), f*-modeling* (*circles*) *and* g*-modeling* (*dots*). *Right panel compares standard deviations of* $\hat{E}\{\theta|x\}$, *for Tweedie estimates* (*dots*), f*-modeling* (*dashed curve*) *and* g*-modeling* (*solid curve*); *reversals at far right are computational artifacts.*

as in (3.11). Brown, Greenshtein and Ritov (2013) use various forms of semiparametric f-modeling to improve on (6.3).

The prehistory of empirical Bayes applications notably includes the *missing species problem*; see Section 11.5 of Efron (2010). This has the Poisson form (6.1), but with an inference different than (6.2) as its goal. Fisher, Corbet and Williams (1943) employed parameterized f-modeling as in Section 4, with f the negative binomial family. Section 3.2.1 of Carlin and Louis (2000) follows the same route for improving Robbins' estimator (6.3).

Tweedie's formula (Efron, 2011) extends Robbins-type estimation of $E\{\theta|x\}$ to general exponential families. For the normal case

$$\theta \sim g(\cdot) \quad \text{and} \quad x|\theta \sim \mathcal{N}(\theta, 1), \tag{6.4}$$

Tweedie's formula is

$$\begin{aligned} &E\{\theta|x\} = x + l'(x) \\ &\qquad \text{where } l'(x) = \frac{d}{dx}\log f(x), \end{aligned} \tag{6.5}$$

with $f(x)$ the marginal distribution of X. As in (6.2), the marginal distribution of X determines $E\{\theta|x\}$, without any specific reference to the prior $g(\theta)$.

Given observations X_k from model (6.4),

$$X_k \sim \mathcal{N}(\Theta_k, 1) \quad \text{for } k = 1, 2, \dots, N, \tag{6.6}$$

the empirical Bayes estimation of $E\{\theta|x\}$ is conceptually straightforward: a smooth estimate $\hat{f}(x)$ is obtained from the X_k's, and its logarithm $\hat{l}(x)$ differentiated to give

$$\hat{E}\{\theta|x\} = x + \hat{l}'(x), \tag{6.7}$$

again without explicit reference to the unknown $g(\theta)$. Modeling here is naturally done on the x-scale. [It is not necessary for the X_k's to be independent in (6.6), or (6.1), although dependence decreases the accuracy of $\hat{E}$; see Theorem 8.4 of Efron (2010).]

Figure 7 concerns an application of Tweedie's formula to the *prostate data*, the output of a microarray experiment comparing 52 prostate cancer patients with 50 healthy controls (Efron, 2010, Section 2.1). The genetic activity of $N = 6033$ genes was measured for each man. Two-sample tests comparing patients with controls yielded z-values for each gene, $X_1, X_2, \dots, X_N$, theoretically satisfying

$$X_k \sim \mathcal{N}(0, 1) \tag{6.8}$$

under the null hypothesis that gene k is equally active in both groups. Of course, the experimenters were searching for activity *differences*, which would manifest themselves as unusually large values $|X_k|$. Figure 2.1 of Efron (2010) shows the histogram of the X_k values, looking somewhat like a long-tailed version of a $\mathcal{N}(0, 1)$ density.

The "smooth estimate" $\hat{f}(x)$ needed for Tweedie's formula (6.7) was calculated by Poisson regression, as in (4.3)–(4.7). The 6033 X_k values were put into 193 equally spaced bins, centered at $x_1, x_2, \dots, x_{193}$, chosen as in (2.8) with y_i being the number in bin i. A Poisson generalized linear model (4.3) then gave MLE $\hat{\mathbf{f}} = (\hat{f}_1, \hat{f}_2, \dots, \hat{f}_{193})$. Here the structure matrix $\mathbf{X}$ was the normal spline basis $ns(\mathbf{x}, df = 5)$ augmented with a column of 1's. Finally, the smooth curve $\hat{f}(x)$ was numerically differentiated to give $\hat{l}'(x) = \hat{f}'(x)/\hat{f}(x)$ and $\hat{E} = x + \hat{l}'(x)$.

Tweedie's estimate $\hat{E}\{\theta|x\}$ (6.7) appears as the solid curve in the left panel of Figure 7. It is nearly zero between -2 and 2, indicating that a large majority of genes obey the null hypothesis (6.7) and should be estimated to have $\theta = 0$. Gene 610 had the largest observed z-value, $X_{610} = 5.29$, and corresponding Tweedie estimate 4.09.

For comparison, $\hat{E}\{\theta|x\}$ was recalculated both by f-modeling as in Section 4 and g-modeling as in Section 5 [with discrete sampling distributions (2.4)–(2.6) obtained from $X_k \sim \mathcal{N}(\Theta_k, 1)$, Θ_k being the "true effect size" for gene k]; f-modeling used $\mathbf{X}$ and $\hat{\mathbf{f}}$ as just described, giving $\hat{E}_f = U_r'\hat{\mathbf{f}}/V_r'\hat{\mathbf{f}}$, U_r and V_r as in (4.19), $r = 12$; g-modeling took $\boldsymbol{\theta} = (-3, -2.8, \dots, 3)$ and $Q = (ns(\boldsymbol{\theta}, 5), \mathbf{1})$, yielding $\hat{\mathbf{g}} = \mathbf{g}(\hat{\alpha})$ as the MLE from (5.1)–(5.2). [The R nonlinear maximizer `nlm` was used to find $\hat{\alpha}$; some care was needed in choosing the control parameters of `nlm`. We are paying for the fact that the g-modeling likelihood (5.2) is not an exponential family.] Then the estimated posterior expectation $\hat{E}_g$ was calculated applying Bayes rule with prior $\hat{\mathbf{g}}$. Both $\hat{E}_f$ and $\hat{E}_g$ closely approximated the Tweedie estimate.

Standard deviation estimates for $\hat{E}_f$ [dashed curve, from Theorem 3 with $\hat{\mathbf{f}}$ replacing $\mathbf{f}$ in (4.9)] and $\hat{E}_g$ (solid curve, from Theorem 4) appear in the right panel of Figure 7; f-modeling gives noticeably lower standard deviations for $E\{\theta|x\}$ when $|x|$ is large.

The large dots in the right panel of Figure 7 are bootstrap standard deviations for the Tweedie estimates $\hat{E}\{\theta|x\}$, obtained from $B = 200$ nonparametric bootstrap replications, resampling the $N = 6033$ X_k values. These closely follow the f-modeling standard deviations. In fact, $\hat{E}_f^*$, the bootstrap replications of $\hat{E}_f$, closely matched $\hat{E}^*$ for the corresponding Tweedie estimates on a case-by-case comparison of the 200 simulations. That is, $\hat{E}_f$ is numerically just about the same as the Tweedie estimate, though it is difficult to see analytically why this is the case, comparing formulas (4.16) and (6.7). Notice that the bootstrap results for $\hat{E}_f$ verify the accuracy of the delta-method calculations going into Theorem 3.

Among empirical Bayes techniques, the James–Stein estimator is certainly best known. Its form,

$$\hat{\theta} = \bar{X} + [1 + (N-3)/S](X_k - \bar{X}) \qquad \left[S = \sum_1^N (X_k - \bar{X})^2\right], \tag{6.9}$$

again has the "classic" property of being estimated directly from the marginal distribution on the x-scale, without reference to $g(\theta)$. The simplest application of Tweedie's formula, taking $\mathbf{X}$ in our previous discussion to have rows $(1, x_i, x_i^2)$, leads to formula (6.9); see Section 3 of Efron (2011).

Perhaps the second most familiar empirical Bayes applications relates to Benjamini and Hochberg's (1995) theory of false discovery rates. Here we will focus on the *local false discovery rate* (fdr), which best illustrates the Bayesian connection. We assume that the marginal density of each observation of X_k has the form

$$f(x) = \pi_0\varphi(x) + (1 - \pi_0) f_1(x), \tag{6.10}$$

where π_0 is the prior probability that X_k is null, $\varphi(x)$ is the standard $\mathcal{N}(0, 1)$ density $\exp(-\frac{1}{2}x^2)/\sqrt{2\pi}$, and $f_1(x)$ is an unspecified nonnull density, presumably yielding values farther away from zero than does the null density φ.

Having observed X_k equal to some value x, fdr(x) is the probability that X_k represents a null case (6.8),

$$\text{fdr}(x) = \Pr\{\text{null}|x\} = \pi_0\varphi(x)/f(x), \tag{6.11}$$

the last equality being a statement of Bayes rule. Typically π_0, the prior null probability, is assumed to be near 1, reflecting the usual goal of large-scale testing: to reduce a vast collection of possible cases to a much smaller set of particularly interesting ones. In this case, the *upper false discovery rate*,

$$\text{ufdr}(x) = \varphi(x)/f(x), \tag{6.12}$$

setting $\pi_0 = 1$ in (6.11), is a satisfactory substitute for fdr(x), requiring only the estimation of the marginal density $f(x)$.

Returning to the discrete setting (2.9), suppose we take the parameter of interest $t(\theta)$ to be

$$\mathbf{t} = (0, 0, \dots, 0, 1, 0, \dots, 0)', \tag{6.13}$$

with "1" at the index j_0 having $\theta_{j_0} = 0$ [$j_0 = 16$ in (2.7)]. Then $E\{t(\theta)|x_i\}$ equals fdr(x_i), and we can

TABLE 4

Local false discovery rate estimates for the prostate data; $\widehat{\text{ufdr}}$ *and its standard deviation estimates sdf obtained from f-modeling*; $\widehat{\text{fdr}}$ *and sdg from g-modeling*; *sdf is substantially smaller than sdg*

x	−4	−3	−2	−1	0	1	2	3	4
$\widehat{\text{ufdr}}$	0.060	0.370	0.840	1.030	1.070	1.030	0.860	0.380	0.050
sdf	0.014	0.030	0.034	0.017	0.013	0.021	0.033	0.030	0.009
sdg	0.023	0.065	0.179	0.208	0.200	0.206	0.182	0.068	0.013
$\widehat{\text{fdr}}$	0.050	0.320	0.720	0.880	0.910	0.870	0.730	0.320	0.040

assess the accuracy of a g-model estimate $\widehat{\text{fdr}}(x_i)$ using (5.18), the corollary to Theorem 4.

This was done for the prostate data, with the data binned as in Figure 7, and $Q = (ns(\boldsymbol{\theta}, 5), \mathbf{1})$ as before. Theorem 4 was applied with $\boldsymbol{\theta}$ as in (2.7). The bottom two lines of Table 4 show the results. Even with $N = 6033$ cases, the standard deviations of $\widehat{\text{fdr}}(x)$ are considerable, having coefficients of variation in the 25% range.

F-model estimates of fdr fail here, the bias/variance trade-offs of Table 2 being unfavorable for any choice of r. However, f-modeling is a natural choice for ufdr, where the only task is estimating the marginal density $f(x)$. Doing so using Poisson regression (4.3), with $\mathbf{X} = (ns(\mathbf{x}, 5), \mathbf{1})$, gave the top two lines of Table 4. Now the standard deviations are substantially reduced across the entire x-scale. [The standard deviation of $\widehat{\text{ufdr}}$ can be obtained from Theorem 3, with $\mathbf{U} = \varphi(x_i)\mathbf{1}$ and $\mathbf{V}$ the coordinate vector having 1 in the ith place.]

The top line of Table 4 shows $\widehat{\text{ufdr}}(x)$ exceeding 1 near $x = 0$. This is the penalty for taking $\pi_0 = 1$ in (6.12). Various methods have been used to correct $\widehat{\text{ufdr}}$, the simplest being to divide all of its values by their maximum. This amounts to taking $\hat{\pi}_0 = 1/\text{maximum}$,

$$\hat{\pi}_0 = 1/1.070 = 0.935 \tag{6.14}$$

in Table 4. [The more elaborate f-modeling program `locfdr`, described in Chapter 6 of Efron (2010), gave $\hat{\pi}_0 = 0.932$.] By comparison, the g-model MLE $\hat{\mathbf{g}}$ put probability $\hat{\pi}_0 = 0.852$ on $\theta = 0$.

7. DISCUSSION

The observed data $X_1, X_2, \ldots, X_N$ from the empirical Bayes structure (1.1)–(1.2) arrives on the x scale but the desired Bayesian posterior distribution $g(\theta|x)$ requires computations on the θ scale. This suggests the two contrasting modeling strategies diagrammed in Table 5: modeling on the x scale, "f-modeling," permits the application of direct fitting methods, usually various forms of regression, to the X values, but then pays the price of more intricate and less stable Bayesian computations. We pay the price up front with "g-modeling," where models such as (5.2) require difficult nonconvex maximum likelihood computations, while the subsequent Bayesian computations become straightforward.

The comparative simplicity of model fitting on the x scale begins with the nonparametric case: f-modeling needs only the usual vector of proportions $\hat{\mathbf{f}}$ (3.10), while g-modeling requires Laird's (1978) difficult nonparametric MLE calculations. In general, g-models have a "hidden" quality that puts more strain on parametric assumptions; f-modeling has the advantage of fitting directly to the observed data.

There is a small circle of empirical Bayes situations in which the desired posterior inferences can be expressed as simple functions of $f(x)$, the marginal distribution of the X observations. These are the "classic" situations described in Section 6, and account for the great bulk of empirical Bayes applications. The Bayesian computational difficulties of f-modeling disappear here. Not surprisingly, f-modeling dominates practice within this special circle.

"Bayes rule in terms of f," Section 2, allows us to investigate how well f-modeling performs outside the

TABLE 5

f-modeling permits familiar and straightforward fitting methods on the x scale but then requires more complicated computations for the posterior distribution of θ; *the situation is reversed for g-modeling*

	Model fitting	**Bayesian computations**
f-modeling	direct	indirect
g-modeling	indirect	direct

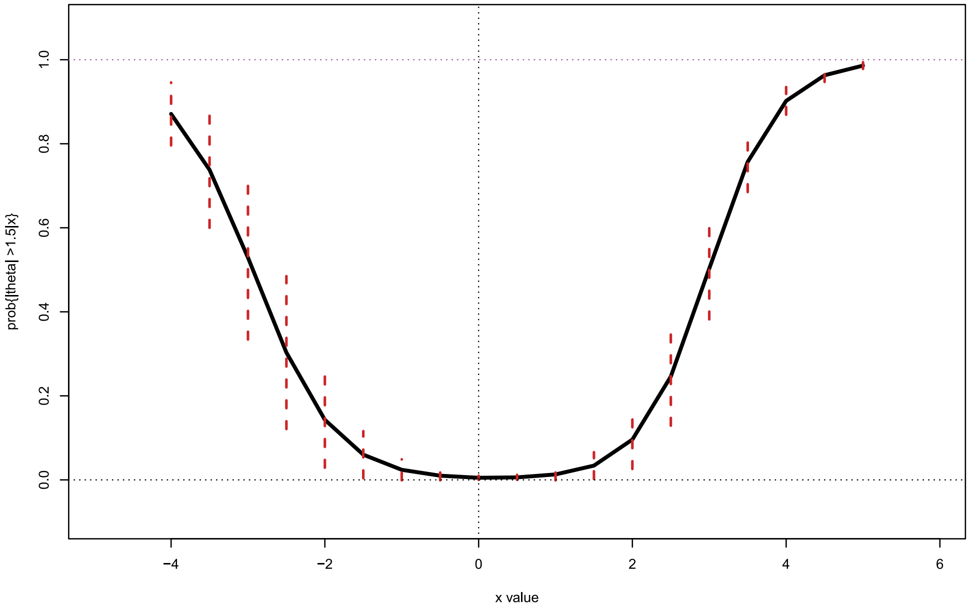

FIG. 8. *g-modeling estimates of* $\Pr\{|\theta| \geq 1.5|x\}$ *for the prostate data. Dashed bars indicate* ± *one standard deviation, from Theorem* 4.

circle. Often not very well seems to be the answer, as seen in the bottom panel of Figure 5, for example. G-modeling comes into its own for more general empirical Bayes inference questions, where the advantages listed in Section 5 count more heavily. Suppose, for instance, we are interested in estimating $\Pr\{|\theta| \geq 1.5|x\}$ for the prostate data. Figure 8 shows the g-model estimates and their standard deviations from Theorem 4, with $Q = ns(\boldsymbol{\theta}, 6)$ as before. Accuracy is only moderate here, but, nonetheless, some useful information has been extracted from the data (while, as usual for problems involving discontinuities on the θ scale, f-modeling is ineffective).

Improved f-modeling strategies may be feasible, perhaps making better use of the kinds of information in Table 2. A reader has pointed out that pseudo-inverses of P other than A (3.1) are available, of the form

$$(P'BP)^{-1}P'B. \tag{7.1}$$

Here the matrix B might be a guess for the inverse covariance matrix of $\hat{\mathbf{f}}$, as motivated by generalized least squares estimation. So far, however, situations like that in Figure 8 seem inappropriate for f-modeling, leaving g-modeling as the only game in town.

Theorems 3 and 4 provide accuracy assessments for f-modeling and g-modeling estimates. These can be dishearteningly broad. In the bottom panel of Figure 5, the "good" choice, g-modeling, would still require more than $N = 20{,}000$ independent observations X_k to get the coefficient of variation down to 0.1 when x exceeds 2. More aggressive g-modeling, reducing the degrees of freedom for Q, improves accuracy, at the risk of increased bias. The theorems act as a reminder that, outside of the small circle of its traditional applications, empirical Bayes estimation has an ill-posed aspect that may call for draconian model choices. [The ultimate choice is to take $g(\theta)$ as known, that is, to be Bayesian rather than empirical Bayesian. In our framework, this amounts to tacitly assuming an enormous amount "N" of relevant past experience.]

Practical applications of empirical Bayes methodology have almost always taken Θ_k and X_k in (1.1)–(1.2) to be real-valued, as in all of our examples. This is not a necessity of the theory (nor of its discrete implementation in Section 2). Modeling difficulties mount up in higher dimensions, and even studies as large as the prostate investigation may not carry enough information for accurate empirical Bayes estimation.

There are not many big surprises in the statistics literature, but empirical Bayes theory, emerging in the 1950s, had one of them: that parallel experimental structures like (1.1)–(1.2) carry within themselves their own Bayesian priors. Essentially, the other $N - 1$ cases furnish the correct "prior" information for analyzing each (Θ_k, X_k) pair. How the statistician extracts that information in an efficient way, an ongoing area of study, has been the subject of this paper.

ACKNOWLEDGMENTS

I am grateful to Omkar Muralidharan, Amir Najmi and Stefan Wager for many helpful discussions. Research supported in part by NIH Grant 8R37 EB002784 and by NSF Grant DMS-12-08787.

REFERENCES

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57** 289–300. MR1325392

BROWN, L. D., GREENSHTEIN, E. and RITOV, Y. (2013). The Poisson compound decision problem revisited. *J. Amer. Statist. Assoc.* **108** 741–749.

BUTUCEA, C. and COMTE, F. (2009). Adaptive estimation of linear functionals in the convolution model and applications. *Bernoulli* **15** 69–98. MR2546799

CARLIN, B. P. and LOUIS, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd ed. *Texts in Statistical Science*. Chapman & Hall, Boca Raton, FL.

CASELLA, G. (1985). An introduction to empirical Bayes data analysis. *Amer. Statist.* **39** 83–87. MR0789118

CAVALIER, L. and HENGARTNER, N. W. (2009). Estimating linear functionals in Poisson mixture models. *J. Nonparametr. Stat.* **21** 713–728. MR2549434

EFRON, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Statist.* **3** 1189–1242. MR0428531

EFRON, B. (2004). The estimation of prediction error: Covariance penalties and cross-validation. *J. Amer. Statist. Assoc.* **99** 619–642. MR2090899

EFRON, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. IMS* **1**. Cambridge Univ. Press, Cambridge. MR2724758

EFRON, B. (2011). Tweedie's formula and selection bias. *J. Amer. Statist. Assoc.* **106** 1602–1614. MR2896860

EFRON, B. and MORRIS, C. (1975). Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.* **70** 311–319.

FISHER, R., CORBET, A. and WILLIAMS, C. (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

HALL, P. and MEISTER, A. (2007). A ridge-parameter approach to deconvolution. *Ann. Statist.* **35** 1535–1558. MR2351096

JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. In *Proc.* 4*th Berkeley Sympos. Math. Statist. and Prob., Vol. I* 361–379. Univ. California Press, Berkeley, CA. MR0133191

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

LAIRD, N. (1978). Nonparametric maximum likelihood estimation of a mixed distribution. *J. Amer. Statist. Assoc.* **73** 805–811. MR0521328

MORRIS, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *J. Amer. Statist. Assoc.* **78** 47–65. MR0696849

MURALIDHARAN, O., NATSOULIS, G., BELL, J., JI, H. and ZHANG, N. R. (2012). Detecting mutations in mixed sample sequencing data using empirical Bayes. *Ann. Appl. Stat.* **6** 1047–1067. MR3012520

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. Univ. California Press, Berkeley and Los Angeles. MR0084919

ZHANG, C. -H. (1997). Empirical Bayes and compound estimation of normal means. *Statist. Sinica* **7** 181–193. MR1441153