Empirical Bayes Methods for Combining Likelihoods

Bradley EFRON

Suppose that several independent experiments are observed, each yielding a likelihood \( L_k(\theta_k) \) for a real-valued parameter of interest \( \theta_k \). For example, \( \theta_1 \) might be the log-odds ratio for a 2 × 2 table relating to the \( k \)th population in a series of medical experiments. This article concerns the following empirical Bayes question: How can we combine all of the likelihoods \( L_k \) to get an interval estimate for any one of the \( \theta_k \)'s, say \( \theta_1 \)? The results are presented in the form of a realistic computational scheme that allows model building and model checking in the spirit of a regression analysis. No special mathematical forms are required for the priors or the likelihoods. This scheme is designed to take advantage of recent methods that produce approximate numerical likelihoods \( L_k(\theta_k) \) even in very complicated situations, with all nuisance parameters eliminated. The empirical Bayes likelihood theory is extended to situations where the \( \theta_k \)'s have a regression structure as well as an empirical Bayes relationship. Most of the discussion is presented in terms of a hierarchical Bayes model and concerns how such a model can be implemented without requiring large amounts of Bayesian input. Frequentist approaches, such as bias correction and robustness, play a central role in the methodology.

KEY WORDS: ABC method; Confidence expectation; Generalized linear mixed models; Hierarchical Bayes; Meta-analysis for likelihoods; Relevance; Special exponential families.

1. INTRODUCTION

A typical statistical analysis blends data from independent experimental units into a single combined inference for a parameter of interest \( \theta \). Empirical Bayes, or hierarchical or meta-analytic analyses, involve a second level of data acquisition. Several independent experiments are observed, each involving many units, but each perhaps having a different value of the parameter \( \theta \). Inferences about one or all of the \( \theta \)'s are then made on the basis of the full two-level compound data set. This article concerns the construction of empirical Bayes interval estimates for the individual parameters \( \theta \), when the observed data consist of a separate likelihood for each case.

Table 1, the ulcer data, exemplifies this situation. Forty-one randomized trials of a new surgical treatment for stomach ulcers were conducted between 1980 and 1989 (Sacks, Chalmers, Blum, Berrier, and Pagano 1990). The \( k \)th experiment’s data are recorded as

\[
(a_k, b_k, c_k, d_k) \quad (k = 1, 2, \ldots, 41),
\]

where \( a_k \) and \( b_k \) are the number of occurrences and nonoccurrences for the Treatment (the new surgery) and \( c_k \) and \( d_k \) are the occurrences and nonoccurrences for Control (an older surgery). Occurrence here refers to an adverse event—recurrent bleeding.

The estimated log odds ratio for experiment \( k \),

\[
\hat{\theta}_k = \log \left( \frac{a_k}{b_k} / \frac{c_k}{d_k} \right),
\]

measures the excess occurrence of Treatment over Control. For example \( \hat{\theta}_{13} = .60 \), suggesting a greater rate of occurrence for the Treatment in the 13th experimental population. But \( \hat{\theta}_8 = -4.17 \) indicates a lesser rate of occurrence in the 8th population. An approximate standard deviation for \( \theta_k \), also appearing in Table 1, is

\[
\hat{\Sigma}D_k = \left\{ \frac{1}{a_k + .5} + \frac{1}{b_k + .5} + \frac{1}{c_k + .5} + \frac{1}{d_k + .5} \right\}^{1/2};
\]

\[
\hat{\Sigma}D_{13} = .61,
\]

for example.

The statistic \( \hat{\theta}_k \) is an estimate of the true log-odds ratio \( \theta_k \) in the \( k \)th experimental population,

\[
\theta_k = \log \left( \frac{P_k(\text{Occurrence}|\text{Treatment})}{P_k(\text{Nonoccurrence}|\text{Treatment})} \div \frac{P_k(\text{Occurrence}|\text{Control})}{P_k(\text{Nonoccurrence}|\text{Control})} \right),
\]

where \( P_k \) indicates probabilities for population \( k \). The data \((a_k, b_k, c_k, d_k)\), considered as a 2 × 2 table with fixed margins, give a conditional likelihood for \( \theta_k \),

\[
L_k(\theta_k) = \left( \frac{a_k + b_k}{a_k} \right) \left( \frac{c_k + d_k}{c_k} \right) \exp\left\{ \theta_k a_k / S_k(\theta_k) \right\}
\]

(Lehmann 1959, sec. 4.6). Here \( S_k(\theta_k) \) is the sum of the numerator in (5) over the allowable choices of \( a_k \) subject to the 2 × 2 table’s marginal constraints.

Figure 1 shows 10 of the 41 likelihoods (5), each normalized so as to integrate to 1 over the range of the graph. It seems clear that the \( \theta_k \) values are not all the same. For instance, \( L_8(\theta_k) \) and \( L_{13}(\theta_{13}) \) barely overlap. On the other hand, the \( \theta_k \) values are not wildly discrepant, most of the 41 likelihood functions being concentrated in the range \( \theta_k \in (-6.3) \).

This article concerns making interval estimates for any one of the parameters, say \( \theta_8 \), on the basis of all the data in Table 1. We could of course use only the data from experiment 8 to form a classical confidence interval for \( \theta_8 \). This amounts to assuming that the other 40 experiments have no relevance to the 8th population. At the other extreme, we could assume that all of the \( \theta_k \) values were equal and form
Our empirical Bayes confidence intervals will be constructed directly from the likelihoods $L_k(\theta_k), k = 1, 2, \ldots, K$, with no other reference to the original data that gave these likelihoods. This can be a big advantage in situations where the individual experiments are more complicated than those in Table 1. Suppose, for instance, that we had observed $d$-dimensional normal vectors $x_k \sim N_d(\mu_k, \Sigma_k)$ independently for $k = 1, 2, \ldots, K$, and that the parameter of interest was $\theta_k =$ second-largest eigenvalue of $\Sigma_k$. Current research has provided several good ways to construct approximate likelihoods $L_{z_k}(\theta_k)$ for $\theta_k$ alone, with all nuisance parameters eliminated (see Barnadoff-Nielsen 1986, Cox and Reid 1987, and Efron 1993). The set of approximate likelihoods, $L_{z_k}(\theta_k), k = 1, 2, \ldots, K$, could serve as the input to our empirical Bayes analysis. In this way empirical Bayes methods can be brought to bear on very complicated situations.

The emphasis here is on a general approach to empirical Bayes confidence intervals that does not require mathematically tractable forms for the likelihoods or for the family of possible prior densities. Our results, which are primarily methodological, are presented in the form of a realistic computational algorithm that allows model building and model checking, in the spirit of a regression analysis. As a price for this degree of generality, all of our methods are approximate. They can be thought of as computer-based generalizations of the analytic results of Carlin and Gelfand (1990, 1991), Laird and Louis (1987), Morris (1983), and other works referred to by those authors. Section 6 extends the empirical Bayes likelihood methods to problems also having a regression structure. This extension gives a very general version of mixed models applying to generalized linear model (GLM) analyses.

The framework for our discussion is the hierarchical Bayes model described in Section 2. Philosophically, this

![Figure 1. Ten of the 41 Individual Likelihoods $L_k(\theta_k)$, (5); $k = 1, 3, 5, 6, 7, 8, 10, 12, 13,$ and 40; Likelihoods Normalized to Satisfy $\int_{-\infty}^{\infty} L_k(\theta_k) d\theta_k = 1$. Stars indicate $L_k(\theta_0)$, which is more negatively located than the other 40 likelihoods; $L_k(\theta_4)$, not shown, is perfectly flat.](image)

<table>
<thead>
<tr>
<th>Experiment</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>$\hat{\theta}$</th>
<th>$\hat{SD}$</th>
<th>Experiment</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>$\hat{\theta}$</th>
<th>$\hat{SD}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>*1</td>
<td>7</td>
<td>8</td>
<td>11</td>
<td>2</td>
<td>-1.64</td>
<td>.86</td>
<td>21</td>
<td>6</td>
<td>34</td>
<td>13</td>
<td>8</td>
<td>-2.22</td>
<td>.61</td>
</tr>
<tr>
<td>*2</td>
<td>8</td>
<td>11</td>
<td>8</td>
<td>6</td>
<td>-2.17</td>
<td>1.14</td>
<td>22</td>
<td>44</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>-1.60</td>
<td>.73</td>
</tr>
<tr>
<td>*3</td>
<td>5</td>
<td>29</td>
<td>4</td>
<td>35</td>
<td>-0.11</td>
<td>.29</td>
<td>23</td>
<td>1</td>
<td>14</td>
<td>54</td>
<td>13</td>
<td>0.57</td>
<td>1.60</td>
</tr>
<tr>
<td>*4</td>
<td>7</td>
<td>29</td>
<td>4</td>
<td>27</td>
<td>-0.68</td>
<td>.66</td>
<td>24</td>
<td>6</td>
<td>15</td>
<td>8</td>
<td>13</td>
<td>-0.43</td>
<td>.64</td>
</tr>
<tr>
<td>*5</td>
<td>3</td>
<td>9</td>
<td>0</td>
<td>12</td>
<td>Inf</td>
<td>1.57</td>
<td>25</td>
<td>0</td>
<td>6</td>
<td>6</td>
<td>0</td>
<td>Inf</td>
<td>2.08</td>
</tr>
<tr>
<td>*6</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>Inf</td>
<td>1.65</td>
<td>26</td>
<td>1</td>
<td>9</td>
<td>5</td>
<td>10</td>
<td>Inf</td>
<td>1.02</td>
</tr>
<tr>
<td>*7</td>
<td>4</td>
<td>13</td>
<td>13</td>
<td>11</td>
<td>-1.35</td>
<td>.68</td>
<td>27</td>
<td>5</td>
<td>12</td>
<td>5</td>
<td>10</td>
<td>-1.88</td>
<td>.73</td>
</tr>
<tr>
<td>*8</td>
<td>1</td>
<td>15</td>
<td>13</td>
<td>3</td>
<td>-4.17</td>
<td>1.04</td>
<td>28</td>
<td>0</td>
<td>10</td>
<td>12</td>
<td>2</td>
<td>Inf</td>
<td>1.60</td>
</tr>
<tr>
<td>*9</td>
<td>3</td>
<td>11</td>
<td>7</td>
<td>15</td>
<td>-0.54</td>
<td>.76</td>
<td>29</td>
<td>22</td>
<td>8</td>
<td>16</td>
<td>16</td>
<td>Inf</td>
<td>1.49</td>
</tr>
<tr>
<td>*10</td>
<td>2</td>
<td>36</td>
<td>12</td>
<td>20</td>
<td>-2.38</td>
<td>.75</td>
<td>30</td>
<td>0</td>
<td>12</td>
<td>11</td>
<td>11</td>
<td>-1.98</td>
<td>.80</td>
</tr>
<tr>
<td>*11</td>
<td>6</td>
<td>6</td>
<td>8</td>
<td>0</td>
<td>Inf</td>
<td>1.35</td>
<td>31</td>
<td>1</td>
<td>14</td>
<td>7</td>
<td>6.</td>
<td>-2.79</td>
<td>1.01</td>
</tr>
<tr>
<td>*12</td>
<td>2</td>
<td>5</td>
<td>7</td>
<td>2</td>
<td>-2.17</td>
<td>1.06</td>
<td>32</td>
<td>8</td>
<td>16</td>
<td>15</td>
<td>12</td>
<td>-0.92</td>
<td>.57</td>
</tr>
<tr>
<td>*13</td>
<td>9</td>
<td>12</td>
<td>7</td>
<td>17</td>
<td>.60</td>
<td>.61</td>
<td>33</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>2</td>
<td>-1.25</td>
<td>.92</td>
</tr>
<tr>
<td>*14</td>
<td>7</td>
<td>14</td>
<td>5</td>
<td>20</td>
<td>.69</td>
<td>.66</td>
<td>34</td>
<td>0</td>
<td>20</td>
<td>5</td>
<td>18</td>
<td>Inf</td>
<td>1.51</td>
</tr>
<tr>
<td>*15</td>
<td>3</td>
<td>22</td>
<td>11</td>
<td>21</td>
<td>-1.35</td>
<td>.68</td>
<td>35</td>
<td>4</td>
<td>13</td>
<td>2</td>
<td>14</td>
<td>.77</td>
<td>.87</td>
</tr>
<tr>
<td>*16</td>
<td>4</td>
<td>7</td>
<td>6</td>
<td>4</td>
<td>-.97</td>
<td>.96</td>
<td>36</td>
<td>10</td>
<td>30</td>
<td>12</td>
<td>8</td>
<td>-1.50</td>
<td>.57</td>
</tr>
<tr>
<td>*17</td>
<td>2</td>
<td>8</td>
<td>8</td>
<td>2</td>
<td>-2.77</td>
<td>1.02</td>
<td>37</td>
<td>3</td>
<td>13</td>
<td>2</td>
<td>14</td>
<td>.48</td>
<td>.91</td>
</tr>
<tr>
<td>*18</td>
<td>1</td>
<td>30</td>
<td>4</td>
<td>23</td>
<td>-1.65</td>
<td>.98</td>
<td>38</td>
<td>4</td>
<td>30</td>
<td>5</td>
<td>14</td>
<td>-0.99</td>
<td>.71</td>
</tr>
<tr>
<td>*19</td>
<td>4</td>
<td>24</td>
<td>15</td>
<td>16</td>
<td>-1.73</td>
<td>.62</td>
<td>39</td>
<td>7</td>
<td>31</td>
<td>15</td>
<td>22</td>
<td>-1.11</td>
<td>.52</td>
</tr>
<tr>
<td>*20</td>
<td>7</td>
<td>36</td>
<td>16</td>
<td>27</td>
<td>-1.11</td>
<td>.51</td>
<td>*40</td>
<td>0</td>
<td>34</td>
<td>0</td>
<td>0</td>
<td>Inf</td>
<td>2.01</td>
</tr>
</tbody>
</table>

NOTE: 41 independent experiments comparing Treatment, a new surgery for stomach ulcer, with Control, an older surgery; (a, b) = (occurrences, nonoccurrences) on Treatment; (c, d) = (occurrences, nonoccurrences) on Control; occurrence refers to an adverse event, recurrent bleeding. If estimated log odds ratio (2), $\hat{SD} = \text{estimated standard deviation for } \hat{\theta}$ (3). Experiments 40 and 41 are not included in the empirical Bayes analysis of Section 3. Stars indicate likelihoods shown in Figure 1. Data from Sacks et al. (1990). Carl Morris, in Part 8 of his Commentary following this article, points out discrepancies between Table 1 and the original data for cities 4, 24, 32, and 33.
is a very attractive framework for the combination of independent likelihoods. In practice, however, it is difficult to apply hierarchical Bayes methods because of the substantial Bayesian inputs that they require.

This article can be thought of as an attempt to make hierarchical Bayes analyses more flexible and practical, and more palatable to frequentists. Section 3 considers a data-based approach to the choice of prior distributions, less rigid than the usual appeal to conjugate families. Section 5, on bias correction, concerns frequentist adjustments to the maximum likelihood estimation of hierarchical parameters. These adjustments correct a simple frequentist approach to better agree with what one would obtain from a full hierarchical Bayes analysis. Section 4 concerns robustness: How we can protect a genuinely unusual case, perhaps the one drug out of many tested that really works, from being overmodified by Bayesian shrinkage to the mean? Section 7 summarizes the various suggestions and results.

2. HIERARCHICAL AND EMPIRICAL BAYES MODELS

Empirical Bayes confidence intervals can be thought of as approximations to a full hierarchical Bayes analysis of a compound data set. That is the point of view taken by Carlin and Gelfand (1990, 1991), Laird and Louis (1987), and this article. This section briefly reviews the connection between hierarchical and empirical Bayes analyses.

Figure 2 diagrams the hierarchical Bayes model as used here. At the most remote level from the data, a hyperprior density \( h(\cdot) \) yields a hyperparameter \( \eta \) that determines the form of a Bayes prior density \( g_\eta(\cdot) \). For example, \( g_\eta(\cdot) \) could correspond to a normal distribution \( N(\eta_1, \eta_2) \), with \( \eta = (\eta_1, \eta_2) \) sampled from the vague hyperprior density \( h(\eta_1, \eta_2)d\eta_1d\eta_2 = d\eta_1d\eta_2/\eta_1\eta_2 > 0 \).

The prior density \( g_\eta(\cdot) \) yields some unobservable real-valued parameters \( \theta_k \),

\[
\theta_0, \theta_1, \theta_2, \ldots, \theta_K \overset{iid}{\sim} g_\eta(\cdot),
\]

with iid meaning independent and identically distributed. (In this section the cases are indexed from zero to \( K \) instead of from 1 to \( K \), with \( \theta_0 \) representing a parameter of particular interest.) Each \( \theta_k \) determines a sampling density \( f_{\theta_k}(\cdot) \) that then yields some observable data \( x_k \),

\[
x_k \sim f_{\theta_k}(\cdot) \quad \text{independently for} \quad k = 0, 1, 2, \ldots, K.
\]

The \( x_k \) can be vector valued, and the densities \( f_{\theta_k}(\cdot) \) can differ in form for the different cases. In fact our methods use only the likelihoods

\[
L_k(\theta_k) \equiv c f_{\theta_k}(x_k),
\]

with \( x_k \) fixed as observed and \( c \) indicating an arbitrary positive constant. In an earlier work (Efron 1993), I showed how likelihoods \( L_k(\theta_k) \) can be accurately approximated even in cases where nuisance parameters affect the distributions of the \( x_k \).

Our task here is to form interval estimates for any one of the \( \theta_k \), say \( \theta_0 \), on the basis of all of the observed data. The Bayes posterior interval for \( \theta_0 \) can be conveniently described by assuming that the “other” data,

\[
x = (x_1, x_2, \ldots, x_K),
\]

are observed before the direct data \( x_0 \) for \( \theta_0 \). This route is illustrated by the dashed lines in Figure 2.

The marginal sampling density, \( d_{\eta}(x) \), is obtained by integrating out the unobserved vector of “other” parameters
\( \theta = (\theta_1, \theta_2, \ldots, \theta_K) \),

\[ d_\eta(x) = \int g_\eta(\theta) f_\theta(x) \, d\theta, \]

(10)

\( g_\eta(\theta) = \prod_{k=1}^K g_{\eta_k}(\theta_k) \) and \( f_\theta(x) = \prod_{k=1}^K f_{\theta_k}(x_k) \). Bayes’s rule gives

\[ h(\eta|x) = c h(\eta) d_\eta(x) \]

(11)
as the conditional density of \( \eta \), given \( x \), with \( c \) a positive constant as before. This provides an induced prior density for \( \theta_0 \):

\[ g_\eta(\theta_0) = \int h(\eta|x) g_\eta(\theta_0) \, d\eta. \]

(12)

A final application of Bayes’s rule gives the posterior density of \( \theta_0 \) based on all of the data:

\[ p_\eta(\theta_0|x_0) = c g_\eta(\theta_0) f_{\theta_0}(x_0) = c g_\eta(\theta_0) L_0(\theta_0). \]

(13)
The generic positive constant \( c \) may differ in the two parts of (13). Hierarchical Bayes interval estimates for \( \theta_0 \) are computed from \( p_\eta(\theta_0|x_0) \). Notice that (13) generalizes the usual form of Bayes’s rule that would apply if \( \eta \) (and so \( g_\eta \)) were known:

\[ p_\eta(\theta_0|x_0) = c g_\eta(\theta_0) L_0(\theta_0). \]

(14)
The generalization consists of averaging \( g_\eta(\theta_0) \) over \( h(\eta|x) \).

The hierarchical Bayes formula (13) neatly separates the information for \( \theta_0 \); the likelihood \( L_0(\theta_0) \) is the direct information about \( \theta_0 \) contained in \( x_0 \); the induced prior density \( g_\eta(\theta_0) \) is the information about \( \theta_0 \) in the other data \( x \). In practice, though, (13) can be difficult to use because of the substantial Bayesian inputs required in Figure 2. Empirical Bayes methods can be thought of as practical approximations to a full hierarchical Bayes analysis.

The maximum likelihood estimate (MLE) of \( \eta \) based on the other data \( x \) is

\[ \hat{\eta} = \arg \max_\eta \{ d_\eta(x) \}. \]

(15)

A familiar empirical Bayes tactic is to replace the induced prior \( g_\eta(\theta_0) \) with the MLE prior \( g_\eta(\theta_0) \) in (13), giving the MLE posterior density:

\[ p_\eta(\theta_0|x_0) = c g_\eta(\theta_0) L_0(\theta_0) \]

(16)

(see Morris 1983). Percentiles of the distribution corresponding to \( p_\eta(\theta_0|x_0) \) provide a simple form of empirical Bayes confidence intervals for \( \theta_0 \). The hyperprior density \( h(\cdot) \) is no longer necessary. This approach and improvements on it are considered in the next three sections. We follow similar ideas of Carlin and Gelfand (1990) and Laird and Louis (1987), particularly in Section 5, which concerns correcting the biases in (16).

As a simple example, suppose that both \( g_\eta(\cdot) \) and \( f_{\theta_k}(\cdot) \) in Figure 2 are normal:

\[ \theta_k \sim N(M, A) \quad \text{and} \quad x_k|\theta_k \sim N(\theta_k, 1) \]

for \( k = 0, 1, 2, \ldots, K \).

Now \( \eta = (M, A) \). Marginally, the components \( x_k \) of \( x \) are independently normal,

\[ d_{M,A}(x_k) \sim N(M, A + 1), \]

(18)

so that the MLE \( \hat{\eta} = (\hat{M}, \hat{A}) \) has

\[ \hat{M} = \bar{x} = \sum_{k=1}^K x_k/K \]

and

\[ \hat{A} = \sum_{k=1}^K (x_k - \bar{x})^2 / K - 1. \]

(19)

\( L_0(\theta_0) \) equals \( ce^{-(\theta_0 - x_0)^2/2} \). We assume that \( \hat{A} \) in (19) is nonnegative.

Formula (16) gives a MLE posterior density for \( \theta_0 \) that is normal,

\[ p_\eta(\theta_0|x_0) \sim N(\bar{x} + (1 - \hat{B})(x_0 - \bar{x}), 1 - \hat{B}), \]

\[ \hat{B} = K / \sum_{k=1}^K (x_k - \bar{x})^2, \]

(20)

so an empirical Bayes one-sided level \( \alpha \) interval for \( \theta_0 \) goes from \( -\infty \) to \( \bar{x} + (1 - \hat{B})(x_0 - \bar{x}) + z(\alpha) \cdot (1 - \hat{B})^{1/2}, \) with \( z(0.95) = 1.645 \) etc.

Formula (20) suggests the empirical Bayes point estimator

\[ \bar{\theta}_0 = \bar{x} + (1 - \hat{B})(x_0 - \bar{x}) \]

(21)

for \( \theta_0 \). This is a slightly different and not as good as the James–Stein estimator,

\[ \bar{\theta}_0 = \bar{x} + (1 - \hat{B})(x_0 - \bar{x}) \left\{ \begin{array}{l} \bar{x} = \frac{\sum_{k=1}^K K}{K + 1} x_k/(K + 1) \\ \hat{B} = (K - 2)/\sum_{k=1}^K (x_k - \bar{x})^2 \end{array} \right\} \]

(22)

see Efron and Morris (1973, lem. 2). The diminished accuracy comes from the fact that (20) does not use \( x_0 \) in the estimation of \( \eta = (M, A) \).

We can imagine estimating \( (M, A) \) as in (19), except using a data set \( x_0, x_1, \ldots, x_{K}, x_{K+1}, \ldots, x_{K}, x_{K+2}, \ldots, x_{K} \), in which \( x_0 \) is repeated \( n_0 \) times. For large \( K \), straightforward calculations show that the choice \( n_0 = 5 \), rather than \( n_0 = 0 \) as in (19), makes the empirical Bayes point estimate (2.16) nearly equal the James–Stein estimate (22). Efron and Morris (1973, sec. 2) derived \( n_0 = 3 \) as the optimal choice for a slightly different estimation problem.

The point here is that \( x_0 \) can reasonably be included in the estimation of \( \eta \) in formula (16). This is done in the examples that follow, though it would also be feasible to carry out the calculations with \( x_0 \) excluded. In these examples \( x \) indicates all of the data, with the cases numbered \( 1, 2, \ldots, K \) as in Section 1.
3. EXPONENTIAL FAMILIES OF PRIOR DENSITIES

The family of possible prior densities \( g_0(\cdot) \) in Figure 2 is an essential ingredient of our empirical Bayes analysis. This section discusses a wide class of priors based on exponential family theory. Exponential family priors are flexible enough to allow a model building and model checking approach to empirical Bayes problems, more in the spirit of a regression analysis than of a traditional conjugate prior Bayes solution. The price for this generality is solutions that are numerical and approximate rather than formulaic and exact.

An exponential family \( \mathcal{F} \) of prior densities \( g_\eta(\theta) \) for a real-valued parameter \( \theta \) can be written as

\[
\mathcal{F} = \{ g_\eta(\theta) = e^{r(\theta) - \phi(\eta)} g_0(\theta), \eta \in A \}. \tag{23}
\]

Here \( \eta \) is a \( p \)-dimensional parameter vector—the natural or canonical vector—and \( r(\theta) \) is the \( p \)-dimensional sufficient vector, depending on \( \theta \). For example, we might take \( p = 3 \) and \( r(\theta) = (\theta, \theta^2, \theta^3)' \). The nonnegative function \( g_0(\theta) \) is the carrier density. The moment-generating function \( \phi(\eta) \) is chosen to make \( g_\eta(\theta) \) integrate to 1 over \( \Theta \), the range of \( \theta \):

\[
\int_\Theta g_\eta(\theta) \, d\theta = \int_\Theta e^{r(\theta) - \phi(\eta)} g_0(\theta) \, d\theta = 1. \tag{24}
\]

The convex set \( A \) is the subset of \( \eta \) vectors for which (24) is possible. All of these are standard definitions for an absolutely continuous exponential family on the real line, following Lehmann (1983, sec. 1.4), though (23) may look peculiar because \( \eta \) is the random variable rather than the parameter.

It is easy to find the marginal MLE \( \hat{\eta} \), (15), in the exponential family \( \mathcal{F} \). For any function \( h(\theta) \), define \( I_\eta(h) \) to be the integral,

\[
I_\eta(h) = \int_\Theta h(\theta) e^{r(\theta) - \phi(\eta)} g_0(\theta) \, d\theta. \tag{25}
\]

If \( h(\theta) \) is a vector or a matrix, then definition (25) applies component-wise. Notice that the integrals in (25) are one-dimensional no matter what the dimensions of \( \eta, t, \) or \( h \). This leads to easy numerical evaluation of \( I_\eta(h) \). Existence of the integral (25) is assumed in what follows, as are other formal properties such as differentiation under the integral sign. Brown (1984, chap. 2) showed why these properties usually hold in exponential families.

The marginal density of \( x_k \), the \( k \)-th component of \( x \) in Figure 2, can be expressed in terms of the likelihood \( L_k(\theta_k) = c f_{0k}(x_k) \),

\[
d_\eta(x_k) = \int_\Theta g_\eta(\theta_k) f_{0k}(x_k) \, d\theta_k = c \int_\Theta g_0(\theta_k) e^{r(\theta_k)} L_k(\theta_k) \, d\theta_k / e^{\phi(\eta)}. \tag{26}
\]

Because

\[
e^{\phi(\eta)} = \int_\Theta g_0(\theta) e^{r(\theta)} \, d\theta = I_\eta(1) \tag{27}
\]

according to (24), formula (26) gives

\[
d_\eta(x_k) = c I_\eta(L_k) / I_\eta(1). \tag{28}
\]

The constant \( c \) depends on \( x_k \) but not on the choices of \( \eta, t(\theta) \) or \( g_0(\theta) \) in (23).

The gradient \( (\partial / \partial \eta) I_\eta(h) \) (i.e. the vector with components \( (\partial / \partial \eta_k) I_\eta(h) \) for \( k = 1, 2, \ldots, p \)) is

\[
\frac{\partial}{\partial \eta} I_\eta(h) = \int_\Theta t(\theta) h(\theta) e^{r(\theta) - \phi(\eta)} g_0(\theta) \, d\theta = I_\eta(th). \tag{29}
\]

Letting \( m_\eta(x_k) \) be the marginal log-likelihood for \( x_k \),

\[
m_\eta(x_k) = \log(d_\eta(x_k)) = \log I_\eta(L_k) - \log I_\eta(1) + \log c, \tag{30}
\]

we obtain the marginal score function \( (\partial / \partial \eta) m_\eta(x_k) \) from (29):

\[
\frac{\partial}{\partial \eta} m_\eta(x_k) = \frac{I_\eta(tL_k)}{I_\eta(L_k)} - \frac{I_\eta(t)}{I_\eta(1)}. \tag{31a}
\]

The total score \( (\partial / \partial \eta) m_\eta(x) \) is the sum of (31a) over the marginally independent components \( x_k \),

\[
\frac{\partial}{\partial \eta} m_\eta(x) = \sum_{k=1}^K \frac{\partial}{\partial \eta} m_\eta(x_k). \tag{31b}
\]

It is easy to compute the marginal MLE \( \hat{\eta} \), (15), from the estimating equation:

\[
\theta = \frac{\partial}{\partial \eta} m_\eta(x) = \sum_{k=1}^K \left[ \frac{I_\eta(tL_k)}{I_\eta(L_k)} - \frac{I_\eta(t)}{I_\eta(1)} \right]. \tag{32}
\]

In the examples that follow, (32) was solved by Newton–Raphson iteration. The second derivative matrix is obtained from (29) and (31):

\[
-\frac{\partial^2}{\partial \eta^2} m_\eta(x) = \sum_{k=1}^K \left\{ \left[ \frac{I_\eta(t^2L_k)}{I_\eta(L_k)} - \left( \frac{I_\eta(t)}{I_\eta(1)} \right)^2 \right] - \left[ \frac{I_\eta(t^2L_k)}{I_\eta(L_k)} - \left( \frac{I_\eta(tL_k)}{I_\eta(L_k)} \right)^2 \right] \right\}. \tag{33}
\]

Here \( (\partial^2 / \partial \eta^2) m_\eta(x) \) is the \( p \times p \) matrix with entries \( (\partial^2 / \partial \eta_k \partial \eta_l) m_\eta(x) \), and for any \( p \) vector \( \mathbf{v} \), the notation \( \mathbf{v}^2 \) indicates the \( p \times p \) matrix \( \mathbf{v} \mathbf{v}^T \). A more intuitive explanation of (32) and (33), relating to the EM algorithm, appears in Remark B.

Newton–Raphson iterates \( \hat{\eta}(0), \hat{\eta}(1), \ldots \) converging to the solution of (32) are obtained from the usual updating equation,

\[
\hat{\eta}(j + 1) = \hat{\eta}(j) - \left[ -\frac{\partial^2}{\partial \eta^2} m_\eta(x) \right]^{-1} \left[ \frac{\partial}{\partial \eta} m_\eta(x) \right]. \tag{34}
\]

The observed information matrix \( -(\partial^2 / \partial \eta^2) m_\eta(x) |_{\hat{\eta}} \) gives an approximate covariance matrix for the MLE \( \hat{\eta} \),

\[
\tilde{\text{cov}}(\hat{\eta}) = \left[ -\frac{\partial^2}{\partial \eta^2} m_\eta(x) |_{\hat{\eta}} \right]^{-1}. \tag{35}
\]
Notice that \( \hat{\eta} \) and \( \overline{\text{cov}}(\hat{\eta}) \) are obtained from the likelihoods \( L_1, L_2, \ldots, L_K \), with no further reference to the original data \( x_1, x_2, \ldots, x_K \). Jackknife estimates of covariance are used as a check on formula (35) in what follows, and in fact give larger estimated variance in our examples.

This program was applied to the ulcer data of Table 1, with cases \( k = 40, 41 \) excluded for reasons discussed later, so \( K = 39 \). The carrier \( g_0(\theta) \) in (23) was taken to the average likelihood,

\[
g_0(\theta) = \frac{1}{K} \sum_{k=1}^{K} L_k(\theta),
\]

with \( \int_{-\infty}^{\infty} L_k(\theta) \, d\theta = 1 \) as in Figure 1. The dotted curve in Figure 3a shows \( g_0 \). Other choices of \( g_0 \), discussed later, made little difference to the estimation of \( g_\eta \).

The results of two different choices of the exponential family \( F, (23), \) are shown in Figure 3: the quadratic model \( t(\theta) = (\theta, \theta^2) \)' and the cubic model \( t(\theta) = (\theta, \theta^2, \theta^3) \)' (We do not need an intercept term in \( t(\theta) \), for example, \((1, \theta, \theta^2)^t\), because constants are absorbed into \( \phi(\eta) \)). The cubic MLE \( g_\eta \) is centrally a little narrower than \( g_\eta \) for the quadratic model, but it is bimodal, with a second mode at the left end of the \( \theta \) scale. The difference looks more dramatic in Figure 3b, which plots the log densities.

The quadratic model forces \( g_0(\theta) \) to be nearly symmetric, whereas the cubic model allows for asymmetry. Does the cubic model provide a significantly better fit to the ulcer data? We can use Wilks’s likelihood ratio criterion to answer this question. Let \( I_{\eta(c)}(h) \) and \( I_{\eta(q)}(h) \) indicate (25) for the cubic and quadratic MLE models. According to (28), Wilks’s statistic is

\[
W = 2 \log \frac{d_{\eta(c)}(x)}{d_{\eta(q)}(x)} = 2 \sum_{k} \log \left( \frac{I_{\eta(c)}(L_k) \cdot I_{\eta(q)}(1)}{I_{\eta(c)}(1) \cdot I_{\eta(q)}(L_k)} \right).
\]

We can compare \( W \) to a \( \chi^2 \) distribution to test for a significantly better fit. In this case the computed value \( W = .64 \) indicates no improvement in going from the quadratic to the cubic model. Other possible improvements on the quadratic model fared no better.

All these calculations excluded cases \( k = 40, 41 \) in Table 1. \( L_{41}(\theta_{41}) \) is perfectly flat, which drops it out of the formulas whether or not we intend to do so. Figure 1 shows that \( L_{40}(\theta_{40}) \) is located in the far negative reaches of the \( \theta \) axis. When case \( k = 40 \) was included in the calculations, \( W \) significantly favored the cubic model over the quadratic. But looking at the individual summands on the right side of (37), case \( k = 40 \) by itself accounted for almost all of \( W \)'s significance. It was on these grounds that case 40 was removed from the analysis as an overly influential outlier. In general, it is a good idea to consider the components of \( W \) rather than just \( W \) alone.

The MLE quadratic prior density \( g_0(\theta) \) has expectation \(-1.22 \) and standard deviation 1.19. How accurate are these values? Jackknife standard errors were computed by removing each case \( k = 1, 2, \ldots, 39 \) in turn and recalculating \( g_\eta \) (which amounts to thinking of the cases \( x_k \) as being randomly sampled from some superpopulation):

\[
\text{Expectation} = -1.22 \pm .26 \\
\text{Standard deviation} = 1.19 \pm .31.
\]

The \pm numbers are one jackknife standard error. These nonparametric jackknife standard errors were larger than those obtained parametrically from (35). In doing the jackknife calculations, it was possible to remove the deleted case from the definition of \( g_0(\theta) \) in (36) and also remove it from the MLE equation (32), thereby accounting for the data-based choice of carrier. But doing so made very little difference here.

The choice of the average likelihood (36) for the carrier is convenient but not at all necessary. Smoother versions of \( g_0(\theta) \), obtained by Gaussian smoothing of (36), were tried on the ulcer data. These gave much different carriers \( g_0(\theta) \) but almost identical estimates of the prior density \( g_\eta(\theta) \); see Remark A. Both here and in the toxoplasmosis example of Section 6, the choice of \( g_0 \) seemed to be the least sensitive aspect of the prior fitting process. A flat density over
the range of \( \theta \) gave almost the same results, but of course this would not always be true. Recent work by the author extends formula (35) to include a data-based choice of the carrier, but such formulas can be avoided by the use of the jackknife as in (38).

In the normal-normal situation of (17), the average likelihood \( g_0(\theta) \) approaches a \( N(M, A + 2) \) distribution as \( K \to \infty \), considerably more broadly supported than \( g_0(\theta) \sim N(M, A) \), the prior density that we are trying to estimate. This wider support will usually be the case, making (36) at least a sufficiently broad choice for the carrier.

**Remark A.** Returning to Figure 2, suppose that \( \theta_1, \theta_2, \ldots, \theta_K \) (but not \( \theta_0 \)) are observed without error, so \( L_k(\theta_k) \) is a delta function at \( \theta_k \). In this case the average likelihood (36) is equivalent to the empirical distribution on \( \{\theta_1, \theta_2, \ldots, \theta_K\} \). Then it is easy to show that the estimated posterior density \( p_\eta(\theta_0|x_0) \), (16), is the discrete distribution putting probability \( L_0(\theta_k)/\sum_{j=1}^K L_0(\theta_j) \) on \( \theta_k \). This is a reasonable estimator, but we might prefer a smoother answer. Smoothness could be achieved by convolving (36) with a normal kernel, as mentioned previously.

**Remark B.** Formulas (32) and (33) can be rewritten in terms of conditional expectations and covariances:

\[
\frac{\partial}{\partial \eta} m_\eta(x) = \sum_{k=1}^K \left[ E_\eta\{t(\theta_k)|x_k\} - E_\eta\{t(\theta_k)\} \right]
\]

and

\[
-\frac{\partial^2}{\partial \eta^2} m_\eta(x) = \sum_{k=1}^K \left[ \text{cov}_\eta\{t(\theta_k)|x_k\} - \text{cov}_\eta\{t(\theta_k)\} \right].
\]

These expressions are related to the EM algorithm and the missing-data principle, the missing data in this case being the unobserved parameters \( \theta_1, \theta_2, \ldots, \theta_K \) (see Little and Rubin 1987, sec. 7.2, and Tanner 1991).

4. **RELEVANCE**

A crucial issue in empirical Bayes applications is the relevance of the “other” data \( x \) in Figure 2 to the particular parameter of interest \( \theta_0 \). Empirical Bayes methods usually pull extreme results toward the group mean. If \( \theta_0 \) represents something desirable like a cure rate, then the relevance of the other data may suddenly seem debatable. This section concerns a scheme for assessing the relevance of the other data to \( \theta_0 \), allowing \( \theta_0 \) to opt out of the empirical Bayes scheme under certain circumstances. This idea is related to the limited translation estimator of Efron and Morris (1971, 1972).

It is easy to modify the hierarchical Bayes model to allow for exceptional cases. We assume that a random choice “C” has been made between two different prior densities for \( \theta_0 \): \( C = A \) with probability \( h_A \), “A” meaning that Figure 2 applies as shown; but \( C = B \) with probability \( h_B = 1 - h_A \), “B” meaning that \( \theta_0 \) is drawn from a different density than \( g_\eta \), say \( g_B \). Typically, \( g_B \) would be a vague, uninformative prior density, not containing much information about \( \theta_0 \). The choice \( C \) is assumed to be independent of all random mechanisms in Figure 2. This amounts to allocating a priori probability \( h_B \) that the information in \( x \) is irrelevant to the estimation of \( \theta_0 \). The statistician must choose \( h_B \), perhaps using a conventional value like .05, but this is the only non-data-based element required in the analysis.

Having observed \( x \), there are two possible prior densities for \( \theta_0 \):

\[
\theta_0 \sim g_A \text{ with probability } h_A,
\]

or

\[
\theta_0 \sim g_B \text{ with probability } h_B = 1 - h_A, \tag{41}
\]

where we have renamed \( g_\eta(\theta_0) \), (12), as \( g_A(\theta_0) \) for convenient notation. These give \( x \) the marginal densities \( d_A(x_0) = \int \int g_A(\theta_0) f_{\theta_0}(x_0) d\theta_0 \) or \( d_B(x_0) = \int \int g_B(\theta_0) f_{\theta_0}(x_0) d\theta_0 \), with the overall marginal density being

\[
d(x_0) = h_A d_A(x_0) + h_B d_B(x_0). \tag{42}
\]

After \( x_0 \) is observed, the a posteriori probabilities for the two choices in (41) are

\[
h_A(x_0) = h_A d_A(x_0)/d(x_0)
\]

and

\[
h_B(x_0) = h_B d_B(x_0)/d(x_0). \tag{43}
\]

The overall posterior density for \( \theta_0 \) is

\[
h_A(x_0)p_A(\theta_0|x_0) + h_B(x_0)p_B(\theta_0|x_0), \tag{44}
\]

where

\[
p_C(\theta_0|x_0) = g_C(\theta_0)f_{\theta_0}(x_0)/d_C(x_0) \tag{45}
\]

for \( C \) equal \( A \) or \( B \).

There are two extreme cases of (44). If \( h_A(x_0) = 1 \), then we are certain that \( x \) is relevant to the estimation of \( \theta_0 \). In this case the a posteriori density for \( \theta_0 \) is \( p_A(\theta_0|x_0) = p_X(\theta_0|x_0) \), (13). If \( h_A(x_0) = 0 \), then \( x \) is completely irrelevant, and the a posteriori density is \( p_B(\theta_0|x_0) \). The example that follows takes \( p_B(\theta_0|x_0) \) to be the confidence density, meaning that by definition the a posteriori intervals based on \( p_B(\theta_0|x_0) \) coincide with confidence intervals for \( \theta_0 \) based only on \( x_0 \); see Remark D.

Bayes rule gives

\[
\frac{h_A(x_0)}{h_B(x_0)} = \frac{h_A}{h_B} R(x_0),
\]

where

\[
R(x_0) = \frac{d_A(x_0)}{d_B(x_0)} = \frac{\int g_A(\theta_0)L_0(\theta_0) d\theta_0}{\int g_B(\theta_0)L_0(\theta_0) d\theta_0}. \tag{46}
\]

At this point it looks like we could use (46) to evaluate \( h_A(x_0) \) and \( h_B(x_0) \), and then take (44) as our overall a posteriori density for \( \theta_0 \), an appealing compromise between
fully including or fully excluding $\theta_0$ from the empirical Bayes inference scheme. But there is a practical difficulty. Uninformative priors such as $g_B(\theta_0)$ are usually defined only up to an unspecified multiplicative constant. This produces a similar ambiguity in $R(x_0)$, say

$$R(x_0) = cr(x_0),$$

(47)

where $r(x_0)$ is a known function but $c$ is an unknown positive constant. What follows is a data-based scheme for evaluating $c$.

Notice that

$$\int h_B(x_0) d(x_0) dx_0 = h_B,$$

(48)

according to (43). This just says that the marginal expectation of $h_B(x_0)$ is $h_B$. Solving (46), (47) gives

$$h_B(x_0) = \frac{1}{1 + \frac{h_A}{h_B} cr(x_0)}.$$

(49)

We can use (48) and (49) to define a method-of-moments estimate $\hat{c}$ for $c$:

$$\frac{1}{K} \sum_{k=1}^{K} \frac{1}{1 + \frac{h_A}{h_B} \hat{c}r(x_k)} = h_B.$$

(50)

Having determined $\hat{c}$, we can now use (46) and (47) to evaluate the compromise a posteriori density (44) for $\theta_0$.

This scheme was applied to the estimation of $\theta_0$ for the ulcer data. The density $g_A$ in (41) was a bias-corrected version of the quadratic MLE $g_{\hat{A}}$ shown in Figure 3 ($g_A = g_{\hat{A}}^1$ of Fig. 6 in the next section). The uninformative prior $g_B$ was taken to be

$$g_B(\theta_0) = cc^{16\theta_8},$$

(51)

for reasons discussed in Remarks C and D.

Figure 4 shows the percentiles of the overall posterior distribution, density (44), as a function of $h_A$, the prior probability that $\theta_8$ should be included in the empirical Bayes scheme. The central .90 interval with $h_A = .95$ is $(-5.26, -1.787)$, compared to $(-5.10, -1.76)$ for $h_A = 1.00$, the full empirical Bayes inference. The likelihood ratio $R(x_0) = d_A(x_0)/d_B(x_0)$ was estimated as .47 from (47) and (50), so $h_A = .95$ results in a posteriori probability $h_A(x_0) = .90$. As a point of comparison, the same computation gives $R(x) = .077$ and $h_A(x_0) = .59$ for $x_0$ representing experiment 40, the excluded case in Table 1.

Remark C. If we choose the uninformative prior $g_B$ correctly, then the percentiles of the posterior density $p_B(\theta_0|x_0) = cg_B(\theta_0)L_0(\theta_0)$ will nearly equal the confidence limits for $\theta_0$ based on $x_0$. The prior density (51) is correct in this sense for the likelihood $L_0(\theta_8)$ obtained by conditioning the eighth $2 \times 2$ ulcer data table on its marginals. This means that the percentiles at $h_A = 0$ in Figure 4 are nearly the standard confidence limits for $\theta_8$ based on $(a_8, b_8, c_8, d_8)$.

Remark D. Formula (51) is a Welch–Peers prior density, one that gives a posteriori intervals agreeing with ordinary confidence intervals to a second order of approximation. Of course, simpler choices, such as $g_B(\theta_0) = constant$, might be used instead. Earlier work (Efron 1993, sec. 6) gives a straightforward numerical algorithm for calculating Welch–Peers densities in general exponential family situations. Formula (51) was actually obtained from the doubly adjusted ABC likelihood (6.26) of that paper.

5. BIASED CORRECTION

The MLE prior density $g_{\hat{A}}(\theta_0)$ tends to be lighter-tailed than the induced prior $g_A(\theta_0)$, (12). This bias makes interval estimates based on the MLE posterior density narrower than they should be according to the corresponding hierarchical Bayes analysis. Bias correction has become a major theme of recent empirical Bayes literature (see, for example, Carlin and Gelfand 1990, 1991, Laird and Louis 1987, and Morris 1983).

This section introduces a new bias correction technique specifically designed for hierarchical Bayes situations. The technique uses confidence interval calculations to approximate the induced prior $g_*(\theta_0)$ that we would have obtained starting from an appropriate uninformative hyperprior density $h(\eta)$. The actual form of $h(\eta)$ is never required here, which is a great practical advantage when dealing with the general exponential families of Section 3.

Suppose that having observed $x$ from the multiparametric family of densities $d_{\eta}(x)$, we wish to make inferences about $\gamma(\eta)$, a real-valued parameter of interest. In what follows, $\gamma(\eta)$ is $g_{\eta}(\theta_0)$ evaluated at a fixed value of $\theta_0$. Let $\gamma_{\alpha}[\alpha]$ be the endpoint of an approximate one-sided level $\alpha$ confidence interval for $\gamma$ having second-order accuracy,

$$\Pr{\gamma < \gamma_{\alpha}[\alpha]} = \alpha + O \left( \frac{1}{K} \right).$$

(52)
Second-order accuracy says that the accuracy of the approximate interval is improving an order of magnitude faster than the usual $1/\sqrt{K}$ rate in the sample size $K$. The ABC intervals (DiCiccio and Efron 1992) are convenient second-order--accurate intervals for the calculations that follow, though other schemes could be used.

The confidence expectation of $\gamma$ given $x$ is defined as

$$\hat{\gamma} = \int_0^1 \gamma_\eta(x) \, dx. \quad (53)$$

This is the expectation of the confidence density, the density obtained by considering the confidence limits to be quantiles; for example, that $\gamma$ exists in $[\gamma_{\eta}(0.01), \gamma_{\eta}(0.99)]$ with probability .01. Efron (1993) shows that the confidence density derived from the ABC limits matches to second order the Bayes a posteriori distribution of $\gamma(\eta)$ given $x$, starting from an uninformative uninformative prior density $h(\eta)$, the Welch–Peers density.

The confidence expectation $\hat{\gamma}$ can be thought of as an adjusted version of the MLE $\hat{\gamma} = \gamma(\hat{\eta})$ that better matches the a posteriori expectation of $\gamma$ given $x$, starting from an uninformative distribution for $\eta$. We will apply this to $\hat{\gamma} = g_{\eta}(\hat{\theta}_0)$ for a fixed value of $\theta_0$, producing the confidence expectation $g_{\eta}^{\dagger}(\theta_0)$ according to definition (53). This is the confidence expectation method of bias correction for $g_{\eta}(\theta_0)$. The appeal of $g_{\eta}^{\dagger}(\theta_0)$ is that it is a direct approximation to the induced prior $g_{\theta}(\theta_0)$, (12), starting from a vague Welch–Peers prior $h(\eta)$, because $g_{\theta}(\theta_0)$ is the a posteriori expectation of $g_{\eta}(\theta_0)$.

The ABC intervals yield a simple approximate form for the confidence expectation (53):

$$\hat{\gamma}^{\dagger} = \hat{\gamma} + \hat{\sigma}(3a - b/\hat{\sigma} + 2c_q). \quad (54)$$

Here $\hat{\sigma}$ is the usual estimate of standard error for $\hat{\gamma}$ and $(a, b, c_q)$ are the three ABC constants described by DiCiccio and Efron (1992), Efron (1993), and Efron and Tibshirani (1993). The numerical calculations needed for $(a, b, c_q)$ are about twice those required to compare $\hat{\sigma}$.

The calculations here were done using abcon, an S program available from statlib@lib.stat.cam.edu by typing the one-line message “send bootstrap funs from S”*. This is a nonparametric version of the ABC method that treats the $K$ likelihoods $L_k(\theta_k)$ as iid observations from a space of random functions.

Here is a simple example where the confidence expectation method can be directly compared to a genuine hierarchical Bayes analysis starting from an uninformative hyperprior. We consider the normal situation (17) and (18) with $K = 20$. The data vector $x$ is taken to be a stylized normal sample,

$$x = c_0 \cdot \left(\Phi^{-1}(0.5)/20, \Phi^{-1}(1.5/20), \ldots, \Phi^{-1}(19.5/20)\right),$$

with the constant $c_0$ chosen to make $\sum_{i=1}^{20} x_i^2/20 = 2$. Then $\bar{M}, \bar{A}$, (19), equals $(0, 1)$, so that the MLE prior $g_{\eta}(\theta_0)$ is the standard normal density $\left(1/\sqrt{2\pi}\right)\exp\left(-.5\theta_0^2\right)$. This is shown as the starred curve on the left half of Figure 5.

The Bayes curve in Figure 5 is the induced prior $g_{\eta}(\theta_0)$ obtained from the uninformative hyperprior $dM dA/(A + 1), A > 0$. It has heavier tails than the MLE $g_{\eta}(\theta_0)$. The confidence expectation density $g_{\eta}^{\dagger}(\theta_0)$, obtained from (54) using abcon, nicely matches the Bayes curve.

Another bias correction technique, called cdf correction here, appears in Figure 5. This is a bootstrap method proposed by Carlin and Gelfand (1990), following suggestions by Cox (1975) and Efron (1987). Cdf correction is a calibration technique introduced to achieve what Morris (1983) calls the empirical Bayes confidence property: for example, that a 95% empirical Bayes interval for $\theta_0$ will contain $\theta_0$ 95% of the time when averaging over the random selection of $\theta_0$ and $(x_0, x)$ in Figure 2. This is a somewhat different criterion than bias correcting $g_{\eta}(\theta_0)$ to match $g_{\eta}(\theta_0)$, and in fact cdf correction does not match the Bayes curve very well in Figure 5.

Details of these calculations were provided in earlier work (Efron 1994, sec. 4), along with another more favorable example of cdf correction. Laird and Louis (1987) Bayesian bootstrap technique was also briefly discussed in that article, where it performed the poorest of the three methods. On the basis of admittedly limited experience, the author slightly prefers the confidence expectation method to cdf correction, and strongly prefers either of these to the Bayesian bootstrap.

The confidence expectation method of bias correction is applied to the ulcer data in Figure 6, again using abcon. Figure 6a shows that $g_{\eta}^{\dagger}(\theta)$ has a substantially heavier left tail than the uncorrected MLE prior $g_{\eta}(\theta)$, but there is not much difference in the right tail. Figure 6b compares the posterior MLE density $p_{\eta}$, (16), for populations 8 and 13 with the corresponding posterior $p_{\eta}^{\dagger}$ replacing $g_{\eta}(\theta)$ with
Bias correction has a large effect on the inference for population 8, but not for population 13. The cdf correction method was also applied to the ulcer data, producing a bias corrected prior density with tails moderately heavier than \( g_\theta^*(\theta) \). Only minor changes resulted in the posterior densities for populations 8 and 13.

Remark E. We could bias correct the posterior density \( p_\theta(x_0) = c g_\theta(\theta_0) L_\theta(\theta_0) \) rather than the prior density \( g_\theta(\theta_0) \). This is the route taken in all of the references. An advantage of the tactic here, bias correcting \( g_\theta(\theta_0) \), is that it need be done only once rather than separately for each \( \theta_k \). In the example of this article, it was numerically more stable than bias correcting \( p_\theta(\theta_0|x_0) \) in cases where \( L_\theta(\theta_0) \) was far removed from \( g_\theta(\theta_0) \).

6. REGRESSION AND EMPIRICAL BAYES

The individual parameters \( \theta_k \) in Figure 2 can be linked to each other by a regression equation, as well as having an empirical Bayes relationship. It is easy to incorporate regression structure into the empirical Bayes methodology of the preceding sections. Doing so results in a very general version of the classical mixed model. The random-effects portion of the mixed model is used to explain the overdispersion in the binomial regression example that follows. There is a large literature concerning normal theory random effects in generalized linear models (see Breslow and Clayton 1993).

We write \( \theta_k \) as the sum of fixed and random effects,

\[
\theta_k = \nu_k + \delta_k, \tag{56}
\]

where the fixed effects \( \nu_k \) depend on known \( q \)-dimensional covariate vectors \( c_k \) and an unknown \( q \)-dimensional parameter vector \( \beta \),

\[
\nu_k = c_k \beta. \tag{57}
\]

The random effects \( \delta_k \) are assumed to be iid draws from an exponential family like (23), with density \( g_\delta(\delta) \) depending on an unknown parameter vector \( \eta \),

\[
g_\eta(\delta) = e^{\eta t(\delta) - \phi(\eta)} g_0(\delta). \tag{58}
\]

As in (25), we define \( I_\eta(h) = \int h(\delta) e^{\eta t(\delta)} g_0(\delta) d\delta \) for any function \( h(\delta) \), the integral being over the range of \( \delta \).

We observe the likelihood functions

\[
L_{x_k}(\theta_k) = c f_{\theta_k}(x_k) \tag{59}
\]

as in (8), and we wish to estimate \((\eta, \beta)\). Define

\[
L_k(\delta_k) = L_{x_k}(\nu_k + \delta_k), \tag{60}
\]

thought of as a function of \( \delta_k \), with \( \nu_k \) and \( x_k \) held fixed. It is easy to verify that the marginal density of \( x_k \) is

\[
d_{\eta, \beta}(x_k) = c I_\eta(L_k)/I_\eta(1), \tag{61}
\]

as in (28). Then (31a,b) follow, giving the maximum likelihood equation with respect to \( \eta \):

\[
0 = \frac{\partial}{\partial \eta} \log d_{\eta, \beta}(x) = \sum_k \left[ \frac{I_\eta(t L_k)}{I_\eta(L_k)} - \frac{I_\eta(t)}{I_\eta(1)} \right]. \tag{62}
\]

Letting

\[
\hat{L}_k \equiv \frac{\partial}{\partial \nu_k} L_{x_k}(\nu_k + \delta_k)
\]

and

\[
\hat{\tilde{L}}_k \equiv \frac{\partial^2}{\partial \nu_k^2} L_{x_k}(\nu_k + \delta_k), \tag{63}
\]

thought of as functions of \( \nu_k \), with \( \delta_k \) and \( x_k \) held fixed, the maximum likelihood equation with respect to \( \beta \) is

\[
0 = \frac{\partial}{\partial \beta} \log d_{\eta, \beta}(x) = \sum_k c_k I_\eta(\hat{L}_k)/I_\eta(L_k). \tag{64}
\]

Together, (62) and (64) determine the maximum likelihood estimates \((\hat{\eta}, \hat{\beta})\), with their solutions being found by
Newton–Raphson iteration. The $p \times p$ second derivative matrix $-(\partial^2/\partial \eta^2) \log d_{\eta, \beta}(x)$ is given by (33). The $q \times q$ matrix $-(\partial^2/\partial \beta^2) \log d_{\eta, \beta}(x)$ is

$$-\frac{\partial^2}{\partial \beta^2} \log d_{\eta, \beta}(x) = \sum_k c_k \left[ \frac{I_{\eta}(L_k)}{I_{\eta}(L_k)} - I_{\eta}(L_k) \right] c_k' \tag{65}$$

and the $p \times q$ off-diagonal second derivative matrix is

$$-\frac{\partial^2}{\partial \eta \partial \beta} \log d_{\eta, \beta}(x) = \sum_k \left[ \frac{I_{\eta}(L_k)}{I_{\eta}(L_k)} - I_{\eta}(L_k) \right] I_{\eta}(L_k) c_k' \tag{66}$$

Table 2 shows the data from a disease prevalence study in 34 cities and El Salvador (Efron 1986). In city $k$, $s_k$ out of $n_k$ subjects tested positive for toxoplasmosis antibodies, $k = 1, 2, \ldots, 34$. We assume a binomial sampling model,

$$s_k \sim \text{Bin}(n_k, \pi_k), \quad \pi_k = \frac{1}{1 + e^{-\theta_k}} \tag{67}$$

where $\theta_k$ is the logit of the true positive rate $\pi_k$ in city $k$. A cubic logistic regression in rainfall was suggested by previous analyses:

$$\theta_k = \beta_0 + \beta_1 r_k + \beta_2 r_k^2 + \beta_3 r_k^3 \tag{68}$$

The solid curve in the left panel of Figure 7 is the MLE cubic logistic regression, transformed back to the probability scale $\pi = 1/(1 + e^{-\theta})$. The cubic model is quite significant but it fails to account for nearly half of the city-to-city variability: the deviances from the cubic fit are about twice as big as they should be under model (67) and (68). This overdispersion can be explained, or at least described, in terms of the random effects model (56)–(58). Roughly speaking, the variance of $\eta(x)$ accounts for about half of the total deviance of the data points from the regression curve.

The mixed model (56) was applied to the toxoplasmosis data, with

$$\nu_k = \beta_0 + \beta_1 r_k + \beta_2 r_k^2 + \beta_3 r_k^3 \tag{69}$$

using the quadratic model $t(\delta) = (\delta, \delta^2)'$ in (58). As a first step, the $\nu_k$ in (69) were fit to the observed proportions $p_k = s_k/n_k$ by ordinary logistic regression, giving an initial estimate $\hat{\beta}(1)$. Notice that this amounts to assuming that the $\delta_k$ all equal zero; that is, $g_{\eta}(\delta)$ puts all of its probability on $\delta = 0$. The initial fitted values $\hat{\nu}_k(1)$ determine likelihoods $L_k(\delta_k) = L_{x_k}(\hat{\nu}_k(1) + \delta_k)$. The carrier density $g_{\eta}(\delta)$ used in (58) was $\sum_k L_k(\delta_k)/34$, with $J_{-5} L_k(\delta_k) d\delta_k = 1$.

Equation (62) was then solved to produce an initial guess $\hat{\nu}(1)$ for $\eta$. The solid curve in the right panel of Figure 7 is $g_{\hat{\eta}}(1)$. Solving (64), with $\eta = \hat{\nu}(1)$, gives a second estimate $\hat{\beta}(2)$ for $\beta$, and so on, iterating toward the MLE ($\hat{\eta}, \hat{\beta}$).
In fact, there was never much change from $(\hat{\eta}^{(1)}, \hat{\beta}^{(1)})$. The dashed lines in Figure 7 show the second estimate, $1/(1 + e^{-\hat{\rho}_0^{(2)}})$ and $g_{\hat{\eta}(2)}$, the latter nearly equaling $g_{\hat{\eta}(1)}$. Using $t(\delta) = (\hat{\delta}, \hat{\delta}^2, \hat{\delta}^3)'$ instead of $t(\delta) = (\delta, \delta^2)'$ did not significantly improve the fit to the data. The bias corrections of Section 5 were small enough to be ignored in this case. Smoother choices of the carrier density $g_0$ had little effect. All in all, the density $g_{\hat{\eta}(1)}$ shown in Figure 7 is a reasonable estimate of the random component in the mixed model (6.1). This density has expectation and standard deviation

\[-.13 \pm .10 \text{ and } .43 \pm .17, \tag{70}\]

as in (38). These conclusions about the $\delta$ distribution can be thought of as an overdispersion analysis of the toxoplasmosis data. The methods used here apply quite generally, though in this case the results are nearly the same as those obtained for the normal theory model of Anderson and Aitkin (1985) or of Breslow and Clayton (1993).

Suppose that we wish to estimate the toxoplasmosis rate in any one city; for instance, city 7. This is similar to the empirical Bayes estimation of $\theta_8$ in Figure 6, except that now an empirical Bayes confidence interval for $\theta_7 = \tilde{v}_7 + \delta_7$ needs to include the variability in $\tilde{v}_7$. The cubic logistic regression estimate $\hat{\beta}^{(1)}$ gave

\[\tilde{v}_7 = .209 \pm .097. \tag{71}\]

The value .097 was obtained by jackknife calculations as in (38), and was almost double the usual standard error that assumes that model (69) is correct. The dashed curve $\hat{g}$ in Figure 8 is the convolution

\[\hat{g} = g_{\hat{\eta}(1)} \oplus N(.209, .097^2). \tag{72}\]

It has standard deviation .53 compared to .43 for $g_{\hat{\eta}(1)}$. The ad hoc estimator (72) is reasonable here, but can be improved by taking into account nonnormality in the likelihood for $\nu_7$.

The solid curve in Figure 8 is an empirical Bayes posterior density for $\theta_7$, obtained by multiplying the binomial likelihood based on $(s_7, n_7) = (2, 12)$ by $\hat{g}$. The resulting .90 interval is

\[\theta_7 \in (-1.54, -.13) \text{ or } \pi_7 = (.18, .47), \tag{73}\]

which, for comparison, does not contain the usual binomial point estimate $\hat{\pi}_7 = 2/12 = .17$.

7. SUMMARY

This article proposes an empirical Bayes methodology for the practical solution of hierarchical Bayes problems. The emphasis is on computer-based analyses that do not require specially chosen mathematical forms for their success. There are four main aspects to our approach:

1. The use of likelihoods $L_k(\theta_k)$ rather than the $x_k$ themselves, as inputs to the empirical Bayes algorithm. This allows the component data sets $x_k$ to be complicated in their own right, each perhaps involving many nuisance parameters in addition to the parameter of interest $\theta_k$. In the ulcer data example of Table 1, each $x_k$ involves two or three nuisance parameters, having to do with the marginal $2 \times 2$ probabilities. The conditional likelihood (5) removes these. Earlier work (Efron 1993) shows how to get approximate likelihoods like (5), with all nuisance parameters removed, in quite general situations.

2. The use of specially designed exponential families of prior densities, (23), rather than conjugate or normal priors. Modern computer technology, combined with the fact that the component likelihoods $L_k(\theta_k)$ are each one-dimensional, makes this numerically practical. Besides being more flexible, the specially designed families allow model building and model checking in the spirit of a regression analysis, as in Figure 3.

3. Allowing an individual component problem to opt out of the empirical Bayes scheme if the relevance of the “other” data appears sufficiently doubtful. This avoids over-shrinkage of particularly interesting results, an unpleasant feature of most empirical/hierarchical Bayes analyses (see Efron 1981, sec. 8). The relevance calculations of Section 4 robustify the empirical Bayes estimation process in a way that allows genuinely unusual cases to stand by themselves. Another way to do this would be by insisting on heavier trials in the prior family (23), but the scheme here is easier to implement and to interpret.

4. A general nonparametric approach to bias correcting the MLE prior density $g_0$. The confidence expectation method of Section 5 fits in well with the goal of emulating a full hierarchical Bayes analysis for Figure 2, under the usual assumption that the hyperprior is uninformative.
Earlier work (Efron 1994) also presents a jackknife implementation of Carlin and Gelfand's cdf correction method that is computationally competitive with the confidence expectation approach. Neither method is yet on a solid theoretical basis, but the specific cases considered here are encouraging.

This methodology is capable of providing empirical Bayes analyses for much more complicated situations than the ulcer or toxoplasmosis examples. For instance, $\theta_k$ might be the proportion of HIV-positive 30-year-old white females in city $k$, while $x_k$ is an extensive epidemiological survey of that city. A reasonable goal of this work is to allow efficient and practical meta-analyses of arbitrarily complicated parallel studies.

[Received November 1993. Revised August 1995.]

REFERENCES


