A SAMPLE REUSE METHOD FOR ACCURATE
PARAMETRIC EMPIRICAL BAYES CONFIDENCE INTERVALS

BY

BRADLEY P. CARLIN and ALAN E. GELFAND

TECHNICAL REPORT NO. 423
NOVEMBER 7, 1989

PREPARED UNDER CONTRACT
N00014-89-J-1627 (NR-042-267)
FOR THE OFFICE OF NAVAL RESEARCH

Reproduction in Whole or in Part is Permitted
for any purpose of the United States Government

Approved for public release; distribution unlimited.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

DTIC ELECTED
DEC 29 1989

89 12 29 051
A SAMPLE REUSE METHOD FOR ACCURATE PARAMETRIC EMPIRICAL BAYES CONFIDENCE INTERVALS

Bradley P. Carlin and Alan E. Gelfand

ABSTRACT

Parametric empirical Bayes methods of point estimation for a vector of unknown parameters date to the landmark paper of James and Stein (1961). The usual approach is to use the mean of the estimated posterior distribution of each parameter, where the estimation of the prior parameters ("hyperparameters") is accomplished through the marginal distribution of the data. While point estimates computed this way usually perform well, interval estimates based on the estimated posterior (called "naive" EB intervals) are not. They fail to account for the variability in the estimation of the hyperparameters, generally resulting in sub-nominal coverage probability in the "EB" sense defined in Morris (1983a).

In this paper we extend the work of Carlin and Gelfand (1989), who proposed a conditional bias correction method for developing EB intervals which corrects the deficiencies in the naive intervals. We show how bias correction can be implemented in general via a Type III parametric bootstrap procedure, a sample reuse method first employed by Laird and Louis (1987). Theoretical and simulation results indicate that intervals which are accurate with respect to nominal coverage ensue. We give two specific applications (to binomial test data and Poisson failure rate data) where we compute simultaneous point and bias corrected interval estimates.

KEY WORDS: Confidence interval; parametric empirical Bayes; parametric bootstrap; bias correction; conditional calibration.
1. INTRODUCTION

In this paper we consider the problem of multiparameter interval estimation in the parametric empirical Bayes (PEB) framework. We consider the familiar exchangeable model, where at the first stage, given θ, the data vectors $Y_i$ are independently distributed as $f_Y(y_i | \theta)$, $i = 1, \ldots, p$. At the second stage, the $\theta_i$ are supposed i.i.d. with distribution $\pi(\theta_i | \eta)$ over $\Theta$, where $\eta$ indexes the family $\pi$. The marginal distribution of $Y$ is $m_Y(y | \eta) = \int f_Y(y_i | \theta_i) \pi(\theta_i | \eta) d\theta_i$, and the conditional independence structure of our model implies that marginally the $Y_i$ are independent. Thus the joint marginal distribution of all the data is $m(y | \eta) = \prod_i m_Y(y_i | \eta)$. The posterior distribution of $\theta$, depends on the data only through $y$, and is denoted by $f(\theta | y, \eta)$, though in the sequel we suppress the subscript on $f$ to simplify the notation.

If $\eta$ were known then point or interval estimates for $\theta$, would be computed via the posterior $f(\theta | y, \eta)$. Generally, however, $\eta$ is unknown. A pure Bayesian approach would place a third stage prior (also known as a “hyperprior”) $\tau(\eta)$ on $\eta$ and then base inference about $\theta$, on the “marginal posterior,”

$$
h(\theta | y) = \int f(\theta | y, \eta) h(\eta | y) d\eta$$

(1.1)

where $h(\eta | y) \propto m(y | \eta) \tau(\eta)$. The hyperprior $\tau(\eta)$ is often given vague specification. Usually computation of (1.1) is an arduous task. In recent years substantial effort has been devoted to developing methodology to calculate the distribution (1.1) and its characteristics. (see for example Naylor and Smith 1982, Tierney and Kadane 1986, Smith et. al. 1985, Smith et. al. 1987, and Gelfand and Smith 1989).

An alternative is the PEB approach, which treats $\eta$ as a fixed unknown, and replaces the integration over $\eta$ in (1.1) with estimation of $\eta$ (usually via maximum likelihood) from the marginal distribution $m(y | \eta)$, obtaining $\hat{\eta} = \hat{\eta}(y)$. Inference is based upon the “estimated posterior,”
Note that (1.2) may be considered an approximation to (1.1) of order $O(p^{-1})$ in the sense that, under mild regularity conditions,

$$E(g(\theta_0) | y) = E(g(\hat{\theta}) | y, \hat{\eta})[1 + O(p^{-1})]$$

(1.3)

(Kass and Steffey, 1988). Expression (1.3) formalizes the fact that PEB estimates are approximately fully Bayesian posterior means.

There is also a substantial amount of literature which demonstrates that, as an estimator of $g(\theta_0)$, $E(g(\theta_0) | y, \hat{\eta})$ often performs well in a decision theoretic sense (see Morris, 1983a for a summary). Unfortunately, interval estimation of $\theta$, through (1.2) has been less successful; such "naive" confidence intervals based on appropriate percentiles of the estimated posterior (either highest posterior density or "equal tail" intervals) are generally too short, and hence fail to attain the nominal coverage probability. The explanation for this problem from a PEB point of view is that we are ignoring the variability in $\hat{\eta}$; from a Bayesian point of view, that we are ignoring the posterior uncertainty about $\eta$. More precisely

$$Var(\theta_0 | y) = E_{\eta | y}[Var(\theta_0 | y, \eta)] + Var_{\eta | y}[E(\theta_0 | y, \eta)],$$

(1.4)

and so the variance estimate based on (1.2), $Var(\theta_0 | y, \hat{\eta})$, will, according to (1.3), only approximate the first term in (1.4). Morris (1983b, 1987) develops improved approximations to (1.4) in special cases while Kass and Steffey (1988) give a general first order approximation.

We propose a more direct attack on the interval estimation problem. We first formalize our objective, nominal conditional or unconditional EB coverage. We then describe how to correct the bias in the naive interval estimate to meet the objective. Most
importantly, we show that for many interesting problems implementation of the bias correction can be accomplished via a sample reuse method. In particular, we employ the Type III parametric bootstrap introduced by Laird and Louis (1987), applying it to two discrete exponential family data sets.

In the EB framework (Morris 1983a,b, Hill 1988) a statistical model is appealingly specified as a collection of joint probability distributions indexed by some parameter,

\[ \mathcal{P} = \{ p_y(y, \theta) \in \Theta \} \tag{1.5} \]

A member of this family is expressible in the sampling form as \( p_y(y, \theta) = f(y | \theta) n(\theta | \eta) \), or in the inferential form as \( p_y(y, \theta) = f(\theta | y, \eta) m(y | \eta) \). Performance of an inference procedure is evaluated over the variability inherent in both \( \theta \) and the data. Thus an unconditional EB confidence set of size \( 1 - \alpha \) for \( g(\theta) \) is a subset \( t(Y) \) of \( \Theta \) such that

\[ \inf_{P} P_{\eta} \{ g(\theta) \in t(Y) \} \geq 1 - \alpha. \tag{1.6} \]

Equation (1.6) has been criticized as being a weak statement. First, we would likely prefer \( P_{\eta} \{ g(\theta) \in t(Y) \} = 1 - \alpha \) over all distributions in \( \mathcal{P} \). Second, we would likely prefer a probability statement which offers conditional calibration given some appropriate data summary (Rubin, 1984). For instance a pure Bayesian interval would be based upon \( \ell(\theta | y) \), and so is conditionally calibrated given all the data. We propose modifying (1.6) to a conditional statement by integrating instead over the distribution \( p(\theta, y | b(y), \eta) \) for a suitable statistic \( b(Y) \). That is, \( t(Y) \) is a conditional \( 1 - \alpha \) EB confidence set for \( g(\theta) \) given \( b(Y) \) if, for each \( b(y) = b \) and \( \eta \),

\[ P_{\eta} \{ g(\theta) \in t(Y) | b(y) = b \} = 1 - \alpha. \tag{1.7} \]

In Section 2, we review the “bias correcting” approach (given in Carlin and Gelfand, 1989) for obtaining intervals achieving this sort of coverage probability using various choices of \( b(Y) \). We then settle on \( b(Y) = \mathcal{L} \) as a natural choice, and elaborate on sam-
ple reuse methods to implement the bias correction. Section 3 discusses the mechanics of hyperparameter estimation and EB point and interval estimation of \( \theta \), in the context of two discrete models, the binomial and the Poisson. For illustration, the analysis of two data sets is given in Section 4. Finally in Section 5 we summarize and comment on some remaining problems.

2. PEB MODELLING AND BIAS CORRECTION

Adopting the EB framework we define the concepts of sufficiency and ancillarity (see Hill, 1988) for the model (1.5).

**Definition 1.** A statistic \( T \) is *sufficient* for \( g(\theta) \) in (1.5) if and only if

- \( T \) is sufficient for \( g(\theta) \) in the posterior family, i.e., \( f(g(\theta) | t, \eta) = f(g(\theta) | y, \eta) \) for all \( \eta \in \mathcal{H} \), and,

- \( T \) is sufficient for \( \eta \) in the marginal family, i.e., \( m(y | t, \eta) = m(y | t) \) for all \( \eta \in \mathcal{H} \).

For example, under the exchangeable structure of Section 1, if \( g(\theta) = \theta \), then usually \( T = (\hat{\eta}, \hat{\eta}) \).

Using this definition it is straightforward to show that \( T \) carries all the information about \( \eta \) in the model with respect to inference about \( g(\theta) \), i.e. we may replace \( Y \) by \( T \) in (1.6) and (1.7). In particular for \( \theta \), we need only study intervals based on \( \mathcal{X} \) and \( \hat{\eta} \).

**Definition 2.** A statistic \( A \) is *ancillary* for \( g(\theta) \) in (1.5) if and only if \( T = (\hat{\eta}, a) \) is minimally sufficient for \( g(\theta) \), where \( \hat{\eta} \) is an estimate of \( \eta \), and \( p(a | \eta) = p(a) \) for all \( \eta \in \mathcal{H} \).

Thus if \( A \) is ancillary, its distribution with respect to the marginal family is free of \( \eta \). In this definition, minimal sufficiency is with respect to (1.5) using Definition 1.
In the sequel we take \( g(\theta) = \theta \), and let \( \mathcal{U} \mid \mathcal{V} \) be a generic notation for the distribution of \( U \) given \( V \). To strengthen the statement (1.6) suppose we replace the integration over \( [0, \hat{\gamma}, \eta] \) in (1.6) with an integration over \( [0, \hat{\gamma}, \eta] \mid b(Y), \eta \). If we take \( b(Y) = \gamma \) then we are in fact integrating over \( [0, \hat{\gamma}, \eta] \), the posterior distribution of \( \theta \).

In the sequel we take \( g(\theta) = \theta \), and let \( \mathcal{U} \mid \mathcal{V} \) be a generic notation for the distribution of \( U \) given \( V \). To strengthen the statement (1.6) suppose we replace the integration over \( [0, \hat{\gamma}, \eta] \) in (1.6) with an integration over \( [0, \hat{\gamma}, \eta] \mid b(Y), \eta \). If we take \( b(Y) = \gamma \) then we are in fact integrating over \( [0, \hat{\gamma}, \eta] \), the posterior distribution of \( \theta \).

In the sequel we take \( g(\theta) = \theta \), and let \( \mathcal{U} \mid \mathcal{V} \) be a generic notation for the distribution of \( U \) given \( V \). To strengthen the statement (1.6) suppose we replace the integration over \( [0, \hat{\gamma}, \eta] \) in (1.6) with an integration over \( [0, \hat{\gamma}, \eta] \mid b(Y), \eta \). If we take \( b(Y) = \gamma \) then we are in fact integrating over \( [0, \hat{\gamma}, \eta] \), the posterior distribution of \( \theta \).

If we instead take \( b(Y) = \gamma \), then our integration is over \( [0, \hat{\gamma}, \eta] = [0, \hat{\gamma}, \eta] \cdot [\hat{\gamma}, \eta] = [0, \hat{\gamma}, \eta] \cdot [\hat{\gamma}, \eta] \cdot [\hat{\gamma}, \eta] \), where we have used Basu's well-known theorem in the last step. Thus if we can find an EB-ancillary statistic, we can develop intervals with EB-coverage conditional on the ancillary merely by integrating first over the full posterior and then over the sampling distribution of our estimator \( \hat{\gamma} \). However in the discrete cases we study, and in fact rather generally, exact ancillary statistics are not available. We have not investigated the notion of approximate ancillaries. If we instead take \( b(Y) = \gamma \), then our integration is over \( [0, \hat{\gamma}, \eta] = [0, \hat{\gamma}, \eta] = [0, \hat{\gamma}, \eta] \cdot [\hat{\gamma}, \eta] \). This choice of \( b \) is appealing since \( \gamma \) is sufficient for \( \theta \) in the posterior family. Additionally, this conditioning is straightforward to implement.

If we denote the \( \alpha \) quantile of the posterior \( f(\theta \mid \mathcal{V}, \eta) \) by \( q(\mathcal{V}, \eta) \), then the so-called equal tail "naive" EB interval can be written as

\[
(q_{\alpha/2}(\mathcal{V}, \hat{\gamma}), q_{1-\alpha/2}(\mathcal{V}, \hat{\gamma}))
\]  

As has already been noted, intervals of this type fail to satisfy (1.6), and their coverage conditional on \( \mathcal{V} \) is also poor. Most of the work in the area of EB confidence intervals has focused on "lengthening" these intervals by using the marginal posterior (1.1), either exactly or approximately. Deely and Lindley (1981) and Rubin (1982) perform the numerical integration and compute \( I_4 \) directly. Morris (1987) advocates use of the member of \( f(\theta \mid \mathcal{V}, \eta) \) whose first two moments agree with the first two (estimated) moments of \( I_4 \). Laird and Louis (1987) suggest the use of a parametric bootstrap sampling method.
(which we discuss in some detail below) to approximate (1.1) where \( h \) is taken to be \( \rho(h | \eta) \), the sampling density of \( \eta \), with the arguments interchanged.

In our view, the weakness of these approaches is that they fail to address the salient issue. the existence and nature of an \( h \) which will be successful in achieving nominal \( \ell \)B coverage. Put simply, they are all concerned with "matching" \( I \) in (1.1), without specifying how to choose \( h \) to begin with. This is understandable, as the whole approach is not directly aimed at attaining a specified level of \( \ell \)B coverage (either conditional or unconditional). Furthermore, while the lengthening of the naive intervals created by the "mixing" in (1.1) is usually desirable, if our estimator \( \hat{\eta} \) is badly biased, the naive intervals can actually turn out to be too long. The real issue is how to correct the bias in (2.1).

A direct approach which would be applied to each tail separately is as follows:

Suppose we define

\[
\begin{align*}
\text{(2.2)} \quad r(\hat{\eta}, \eta, y, \alpha) &= P(\theta_i \leq q_x(y, \eta) | \theta_i \sim f(\theta_i | y, \eta))
\end{align*}
\]

and

\[
\begin{align*}
\text{(2.3)} \quad R(\eta, y, \alpha) &= E_{\eta, \hat{\eta}, y} \{r(\hat{\eta}, \eta, y, \alpha)\}.
\end{align*}
\]

If we then solve

\[
\begin{align*}
\text{(2.4)} \quad R(\eta, y, \alpha') &= \alpha
\end{align*}
\]

for \( \alpha' = \alpha'(\hat{\eta}, y, \alpha) \), we conditionally "correct the bias" in using \( \hat{\eta} \). Of course (2.4) is not solvable as it stands since \( \eta \) is unknown. Using \( \hat{\eta} \), we propose to obtain a bootstrap estimate of the left hand side of (2.4) and thus solve instead for \( \alpha'(\hat{\eta}, y, \alpha) \). We correct both the lower and upper percentiles of our naive interval, obtaining \( \alpha_L ' \) and \( \alpha_U ' \), and then take
Carlin and Gelfand (1989) showed that this bias corrected confidence interval is
unique provided \( \partial r/\partial \alpha \) exists. They further showed that if the distributions
\( f(\theta | \xi, \eta) \) and \( g(\hat{\eta} | \xi, \eta) \) are stochastically ordered in \( \eta \), then conditional bias correction
"works" with respect to (1.7), i.e. \( E_{\gamma, \eta} E_{\hat{\gamma}, \eta} g(\hat{\eta}, \eta, \alpha' | \hat{\eta}, \gamma, \eta) \), the expected conditional
bias corrected tail probability, falls in an interval containing \( \alpha \) (for some \( \eta \), it will be
slightly more than \( \alpha \); for others, slightly less). They also provide simulation support in
several elementary cases. Note that while we describe bias correction of the naive in-
terval (2.1) we could just as well have chosen to bias correct intervals resulting from
(1.1). Such correction could be used to possibly obtain shorter intervals achieving de-
sired coverage levels.

Let us briefly investigate the mechanics of bias correction. If \( \pi(\theta | \eta) \) is chosen as the
standard conjugate prior for \( f(\gamma | \theta) \) then of course the posterior \( f(\theta | \gamma, \eta) \) belongs to the
same standard family, and \( r \) in (2.2) takes the relatively simple form

\[
r(\hat{\eta}, \eta, \gamma, \alpha) = F_{\xi, \eta}^{-1}[F_{\xi, \eta}^{-1}(\alpha)] 
\]  

where \( F \) is the posterior c.d.f. To compute \( R \) in (2.3) necessitates integrating over the
distribution \( g(\hat{\eta} | \gamma, \eta) \). If \( g \) is available in closed form this will require a numerical in-
tegration (either Monte Carlo or other quadrature method). Then we compute the
marginal MLE \( \hat{\eta} \) and solve equation (2.4) at \( \eta = \hat{\eta} \) for \( \alpha' \) via regula falsi. This replace-
ment of an unknown population parameter by its estimate from the marginal empirical
c.d.f. is referred to as a "parametric bootstrap" by Laird and Louis (1987).

Examples where this procedure can be carried out efficiently and easily are discussed
in Carlin and Gelfand (1989). However, in many cases (in particular the discrete settings
in Section 3 a closed form for \( g(\hat{\eta} | \Sigma, \eta) \) is unavailable. In fact, in several elementary conjugate situations, the marginal MLE \( \hat{\eta} \) itself has no closed form and can only be computed numerically. In such cases we estimate \( R(\eta, \Sigma, \alpha) \) through the use of a "Type III parametric bootstrap," a sample reuse method introduced by Laird and Louis (1987) and modified here for the case of EB coverage conditional on \( \Sigma \).

Let us first review the unconditional version. To estimate expectations under the sampling density of \( \hat{\eta}, p(\hat{\eta} | \eta) \), we obtain \( \eta \) from the distribution \( \rho(\cdot | \hat{\eta}) \) as follows. Draw \( \theta_0, ..., \theta_{\rho} \) i.i.d. from \( p(\theta | \hat{\eta}) \), then draw \( \Sigma^* \) independently from \( f(\Sigma | \theta, \eta) \), \( i = 1, ..., \rho \), and finally compute \( \eta^* \) from the "pseudodata" \( \{\Sigma^*\} \) in the same way that \( \hat{\eta} \) was computed from the data \( \{\Sigma\} \). Concisely, we have

\[
\hat{\eta} \rightarrow \{\theta^*_i\} \rightarrow \{\Sigma^*_i\} \rightarrow \eta^* \tag{2.7}
\]

For correction conditional on \( \Sigma \), we modify the Laird and Louis procedure in order to draw observations from \( g \) rather than \( \rho \) by changing (2.7) to

\[
\hat{\eta} \rightarrow \{\hat{\theta}^*_k, k \neq i\} \rightarrow \{\Sigma^*_k, k \neq i\} \rightarrow \eta^* = \eta^*(\Sigma^*_k, \{\Sigma^*_k, k \neq i\}) \tag{2.8}
\]

Repeating this process \( N \) times we obtain \( \eta^*_j \stackrel{\text{id}}{=} g(\cdot | \hat{\eta}, \Sigma), j = 1, ..., N \), and our Type III parametric bootstrap estimate of \( R(\eta, \Sigma, \alpha') \) becomes

\[
\frac{1}{N} \sum_{j=1}^{N} r(\eta^*_j, \hat{\eta}, \Sigma, \alpha') \tag{2.9}
\]

We equate (2.9) to \( \alpha \), and solve for \( \alpha' \) by regular falsi as above. Note that since only the \( \Sigma \) are needed to calculate \( \eta^* \), if \( m(\Sigma | \eta) \) can be sampled from directly we can omit the generation of the \( \theta^*_i \) (see Example 3.2).

The Type III parametric bootstrap can obviously be modified to implement bias correction conditional on \( b(Y) \) other than \( \Sigma \). However the theoretical results below (2.5) are only established given \( \Sigma \).
3. SPECIFIC DISCRETE MODELS

We now apply the preceding methodology to perhaps the two most common discrete models.

Example 3.1: Binomial-beta. Suppose the data vectors $\mathcal{X}$ are simply vectors of zeroes and ones, each element corresponding to a success or failure on the $j^{th}$ independent Bernoulli trial in the $i^{th}$ population, $j = 1, \ldots, n$, $i = 1, \ldots, p$. At the first stage of the hierarchy, we assume the probability of success on any given trial to be the same for all trials within the $i^{th}$ population, but possibly different across populations. By sufficiency we reduce to $X_i = \sum_j X_{i,j}$, and so $X_i | \theta_i \sim \text{Bin}(n_i, \theta_i)$. Under the conjugate prior $\theta_i | \eta_i \sim \text{Beta}(a,b)$, $i = 1, \ldots, p$, where $\eta_i = (a,b)$, $a, b > 0$, the posterior distribution is $f(\theta_i | x_i, a, b) = \text{Beta}(a + x_i, b + n_i - x_i)$, and the marginal distribution is $m(x_i | a, b) = \binom{n_i}{x_i} B(a + x_i, b + n_i - x_i) / B(a, b)$, where $B(\cdot, \cdot)$ represents the beta function. The marginal likelihood,

$$
L(a, b) = m(x | a, b) = \prod_{i=1}^p \binom{n_i}{x_i} B(a + x_i, b + n_i - x_i) / B(a, b),
$$

is maximized via numerical methods to produce the marginal MLE, $\hat{\eta} = (\hat{a}, \hat{b})$. The naive EB confidence interval (2.1) is computed as $(Y_{\hat{a}+b, \hat{b}+n-x}^{1/2}, Y_{\hat{a}+b, \hat{b}+n-x}^{1/2}) / N$, where $Y_{a,b}$ is the c.d.f. of a beta distribution with parameter $c$ and $d$.

To implement the bias correction we note that (2.2) becomes

$$
r(\hat{\eta}, \eta, x, a) = Y_{\hat{a}+b, \hat{b}+n-x}(Y_{\hat{a}+b, \hat{b}+n-x}^{-1}(a))
$$

which is available numerically using a scientific subroutine library. Using the Type III parametric bootstrap procedure (2.9) becomes

$$
\frac{N}{\sum_{j=1}^N \binom{n'-x_j}{x_j} B(a_j + x_j, b_j + n - x_j)^{1/2}}
$$

where $N$ is the number of bootstrap samples.
which we equate to \( \alpha \) and solve for \( \alpha' \). Setting \( \alpha \) to the desired nominal level (e.g., \( \alpha = (.05, .95) \) for 90\% conditional EB coverage), we solve for the corresponding \((\alpha'_1, x'_1)\), and compute our interval for \( \theta \), using (2.5). Note that we bias correct each \( \theta_i \)-interval separately, \( i = 1, \ldots, p \), but of course the correction in each case depends on the data through \((\hat{\alpha}, \hat{b})\). If we desire intervals corrected only for unconditional EB coverage, our bootstrap equation becomes

\[
\sum_{j=1}^{J} Y_j (a + x'_j; b + n_j - x'_j) \frac{(Y_j^{-1} - 1; b_j + n_j - x'_j)}{N} = \alpha. \tag{3.4}
\]

Note that (3.4) differs from (3.3) only in replacing the given value \( x \) by the bootstrapped \( x'_j \)'s. Now we are "averaging over \( X \)" as in (1.6), rather than conditioning on \( X \). We must still bias correct each \( \theta_i \)-interval separately, unless we have a balanced experiment (all \( n_j \) equal). In this case the \( x'_j \)'s are marginally exchangeable, hence so are the \( x'_i \)'s, and so we need only solve (3.4) once for \((\alpha'_1, \alpha'_0)\) before using (2.5).

**Example 3.2: Poisson-gamma.** Suppose we have data of the form \((x_i, t_i), i = 1, \ldots, p\), where the \( x_i \) are observed counts during the time interval \((0, t_i)\). For example, the data might be calls arriving at \( p \) different switchboards in the same county. We assume \( X_i | \theta_i \sim \text{Poisson} (\theta_i t_i), i = 1, \ldots, p \), where the \( t_i \) are known "time exposures." Under the conjugate prior, \( \theta_i | \eta \sim \text{Gamma}(a, b) \) (again \( \eta = (a, b), a, b > 0 \)), the posterior distribution \( \theta_i | x_i \) is \( \text{Gamma}(a + x_i, (t_i + 1/b)^{-1}) \), and the marginal distribution of \( X \) is Negative Binomial, i.e.,

\[
m(x_i | a, b) = \binom{x_i + a - 1}{a - 1} \left( \frac{t_i}{t_i + 1/b} \right)^x \left( \frac{1/b}{t_i + 1/b} \right)^a. \tag{3.5}
\]

If \((\hat{\alpha}, \hat{b})\) maximize the marginal likelihood \((\hat{\alpha} \) and \( \hat{b} \) are not available in closed form; see Section 4), then the naive EB confidence interval for \( \theta \), is
\[
\left( D_{2(i+1/2)}^{-1}(x^2/(2(t_i + 1/\hat{b}_j))) \cdot D_{2(i+1/2)}^{-1}(1 - x^2/(2(t_i + 1/\hat{b}_j))) \right). \tag{3.6}
\]

where \( D \) denotes the c.d.f. of a chi-square distribution with \( k \) (not necessarily integer) degrees of freedom.

In this case (2.2) becomes

\[
D_2(a + x_j)[(t_i + 1/\hat{b}_j) / (t_i + 1/\hat{b})] D_2^{-1}(a). \tag{3.7}
\]

For conditional correction analogous to (3.3), the Type III parametric bootstrap (2.9) becomes

\[
\sum_{j=1}^{N} D_2(a + x_j)[(t_i + 1/\hat{b}_j) / (t_i + 1/\hat{b})] D_2^{-1}(a) / N = \alpha \tag{3.8}
\]

which we equate to \( \alpha \) and solve for \( \alpha' \). For unconditional correction analogous to (3.4) we have the equation

\[
\sum_{j=1}^{N} D_2(a + x_j)[(t_i + 1/\hat{b}_j) / (t_i + 1/\hat{b})] D_2^{-1}(a) / N = \alpha \tag{3.9}
\]

The remark after (2.9) reminds us that in this case we would generate negative binomial \( X_i^* \)'s directly. In addition, if \( t_i = t \) for all \( i \), we need only solve (3.9) once for \( (\alpha', \alpha'' \cdot \cdot \cdot) \) before using (2.5).

In this example if we take the gamma scale hyperparameter \( b \) to be known (say \( b = 1 \) w.l.o.g.), and if we assume \( t_i = t \) (= 1 w.l.c.g.), then the marginal family (3.5) is Negative Binomial \((a, 1/2)\). The method of moments (instead of maximum likelihood) estimator of \( a \) is \( \hat{a} = \bar{X} = \sum_{i=1}^{n} X_i/p \) and the distribution of \( \hat{a} \mid X \) a follows from writing \( \hat{a} = (W + X)/p \) where \( W = \sum_{i=1}^{n} X_i \sim \text{Negative Binomial} (a(p - 1), 1/2) \). Thus we can integrate (2.4) directly, avoiding the Type III resampling algorithm.
We illustrate the implementation of the bias corrected naive EiB confidence interval approach with two discrete data sets. Results of several simulation studies evaluating coverage probabilities and interval lengths are discussed in Carlin (1989).

The data set in Table 1 comes from Burton and Turvey (1988), and gives the results of a psychological evaluation of six independent subjects. Each subject was given three hollow wooden balls, two of which contained a small pyramid and the third either a hemisphere, a block, a cylinder, or a cone. On each of the four resulting trials, the subject had to guess which ball was not like the others simply by shaking, turning, etc. Table 1 gives the results of this (balanced) experiment, where $x_i$ is the number of questions answered *incorrectly* by subject $i$, and $r$, is the raw failure rate, which of course is also the usual classical point estimate of $\theta_i$, the true failure rate for subject $i$.

In terms of modelling this experiment, our binomial-beta model seems natural. At the first stage (given $\theta_i$), a subject's four responses could reasonably be assumed to be i.i.d. Bernoulli trials, and the beta provides a broad choice for the second stage distribution of the $(\theta_i)$. In fact, since our prior belief is that the questions are relatively easy, the simpler family Beta$(1, b)$, $b > 1$ seems adequate. Finally this data set benefits substantially from empirical Bayes modelling, since the small amount of information on each subject severely limits frequentist inference.

The results of our analysis are given in Table 2. Since we are in a balanced (i.i.d.) case, the results for subjects 1 and 2 (who both answered two questions incorrectly) are identical, as are those for subjects 4, 5 and 6 (all of whom made no mistakes). Note the familiar shrinking of the classical point estimates toward the overall proportion of
questions answered incorrectly ($5.24 = .208$) by the EB point estimator. Classical confidence intervals (based on Fisher's exact test statistic for $\theta$) and naive EB intervals are shown, along with versions that are bias corrected unconditionally (via (3.41)) and conditionally (via (3.3)) for nominal coverage $\gamma = .90$. We chose $N = 1000$ bootstrap reps in solving these two equations, using FORTRAN augmented by the IMSL subroutine library. One can see that both types of bias corrected intervals have lengths that are between those of the classical and naive EB methods. However, since nominal conditional coverage implies nominal unconditional coverage, neither bias corrected interval can be uniformly shorter. In particular we see that the conditional intervals are shorter when $X = 2$, but that the unconditional intervals are shorter when $X = 0$.

The data presented in Table 3 record numbers of pump failures, $X_i$, observed in thousands of hours, $t_i$, for $p = 10$ different systems of a certain nuclear power plant. The observations are listed in increasing order of raw failure rate $r_i = X_i/t_i$, which again is the classical point estimate of the true failure rate $\theta$, for the $i^{th}$ system. This data originally appeared in Worledge, Stringham, and McClymont (1982), and was subjected to an empirical Bayes analysis by Gaver and O’Muircheartaigh (1987).

(Insert Table 3 about here)

Our approach is that of Example 3.2 above, using the conjugate Gamma($a, b$) prior for $\theta$. Gaver and O’Muircheartaigh also explore this approach, but after computing the estimated posterior (also gamma, of course) in the usual way, they obtain an approximate EB confidence interval for $\theta$, by assuming that the posterior distribution of $\xi = \log(\theta)$ is approximately normal. Thus their EB point estimate for $\theta$, is $\exp(\hat{\mu})$, and their naive 90% EB confidence interval for $\theta$, is given by

\[
\left( \exp(\hat{\mu} - 1.645\hat{\sigma}), \exp(\hat{\mu} + 1.645\hat{\sigma}) \right),
\]

\[\tag{4.1}\]

14
where \( \hat{\mu} \) and \( \hat{\sigma} \) are the mean and standard deviation, respectively, of the (log-gamma) estimated posterior for \( \varepsilon \). (Actually, the authors' concern about conjugate priors' overshrinkage of outliers leads them to prefer a heavier-tailed log-Student's t prior on \( \varepsilon \). However, they conclude that this assumption does not greatly affect the results. To make a fair comparison, we will compare our intervals only with their gamma-based intervals.)

Solving (3.8) and (3.9) involves computing \( N \) MLE vectors \((\hat{a}, \hat{b})\), one for each bootstrapped pseudodata sample. In order to expedite the maximization Gaver and O'Muircheartaigh suggest using as starting values crude moments estimators obtained by equating the first two sample moments of the crude rates, \( \bar{r} \) and \( \bar{s} \), to the corresponding moments in the marginal family, namely, \( E(r) = \bar{E}(X)/t = ab \) and \( \bar{V}ar(r) = \bar{V}ar(X)/t^2 = ab/t + ab^2 \). This results in

\[
\hat{a}_{MOM} = \bar{r}^2/\{s^2 - \bar{r}^{-1}\} \quad (4.2)
\]

and

\[
\hat{b}_{MOM} = [s^2 - \bar{r}^{-1}]/\bar{r} \quad (4.3)
\]

where \( \bar{r}^{-1} = \sum t^{-1}/p \). (These estimators do not exist if \( s^2 \leq \bar{r}^{-1} \), whence any other reasonable values, possibly the "parent" values \( (\hat{a}, \hat{b}) \), can be used.) Our algorithm begins with Newton-Raphson. If it diverges, we use a local grid search to find a better (higher likelihood) place to restart. We iterate grid search and Newton-Raphson up to twenty times, finally giving up and regenerating new pseudodata if the algorithm still fails. We were able to keep the "failure rate" under 1% using this algorithm.

Gaver and O'Muircheartaigh felt that the intervals (4.1) were likely to be too short due to hyperparameter estimation. Their ad hoc remedy was to compute approximate joint 95% confidence regions for their computed values of \( (\hat{a} = .8223, \hat{b} = .7943) \) using a chi-squared likelihood ratio technique. They then searched over all \( (\hat{a}, \hat{b}) \) in this re-
region, taking the largest value of \( \exp(\hat{\mu} + 1.645\hat{\sigma}) \) obtained as their “corrected” upper confidence limit for \( \theta \). (A similar procedure could have been undertaken for the lower limit, but was not since the authors’ interest was, understandably, only in how large the failure rates could be.) While this process certainly does increase the upper confidence limit, it is not clear what confidence level to attach to it.

(Insert Table 4 about here)

Table 4 gives the results for observations 1, 5, 6, and 10 using the classical interval, 
\[
\left( \frac{1}{2} \cdot D_2^*; \left( \alpha/2 \right), \frac{1}{2} \cdot D_2^*; n(1 - \alpha/2) \right),
\]
applied to the raw rates, the naive and corrected Gaver and O’Muircheartaigh (G & O) methods, the naive EB interval (2.1), and the conditional and unconditional bias corrected methods (3.8) and (3.9) above. Bias corrected percentile \( (\alpha') \) values used in these methods are shown in parentheses below the corresponding interval endpoint.

Some conclusions are: while the classical point estimates seem all right, the corresponding interval estimates are very wide. The naive G & O point estimates have been uniformly shrunk closer to zero than the regular EB point estimates, while the corresponding upper confidence limits are uniformly further from zero. The naive G & O interval estimates are not necessarily too short; in the case of the smallest observation, the naive G & O upper limit is already larger than either of the bias corrected upper limits! The “corrected” G & O upper limit is always much larger than that of any of the other EB methods, reflecting the substantial conservatism embodied in this procedure. Note that more bias correcting (more extreme values of \( \alpha' \), hence longer confidence intervals) is present for cases having shorter history (smaller \( t \)) -- observation #5, for example. This jibes with our intuition about EB point estimation: that cases with less information have more uncertainty associated with them and possibly exhibit more extreme shrinkage patterns. It also appears that conditioning on the value of a more highly variable \( X \) (like observation #5) results in a longer interval than would have been obtained had only unconditional coverage been sought.
5. CONCLUSION

In this paper we have described a sample reuse method (the Type III parametric bootstrap) to correct the bias in naive empirical Bayes confidence intervals when tractable distribution theory is unavailable. Through data examples we have shown that this method is easily implemented yielding intervals which retain much of the intuitive appeal associated with PEB and shrinkage estimation. Future effort is directed at more general applications. For example, the conjugate prior assumption may be dropped, using numerical methods to evaluate (2.6) in the absence of a convenient form for the posterior. Additionally we might choose to bias correct intervals based on the marginal posterior (1.1), rather than the naive intervals based on the estimated posterior (1.2). Beginning with a richer class of intervals could lead to shorter corrected intervals achieving nominal coverage. Finally, efforts to unite the PEB and hierarchical Bayesian literature, as in Section 1 of this paper as well as Kass and Steffy (1988), continue to be important.

REFERENCES


### TABLE 1. Psychological Test Data

<table>
<thead>
<tr>
<th>Subject</th>
<th>n</th>
<th>X</th>
<th>r = \hat{\theta}_{Chown}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>2</td>
<td>.50</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>2</td>
<td>.50</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>1</td>
<td>.25</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>0</td>
<td>.00</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>0</td>
<td>.00</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>0</td>
<td>.00</td>
</tr>
</tbody>
</table>

Source: Burton and Turvey (1988)

### TABLE 2. Psychological Test Data Analysis

<table>
<thead>
<tr>
<th>Subject</th>
<th>90% Classical CI</th>
<th>EB</th>
<th>90% Naive EB CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower</td>
<td>Upper</td>
<td>Length</td>
</tr>
<tr>
<td>1, 2</td>
<td>0.500</td>
<td>0.690</td>
<td>0.582</td>
</tr>
<tr>
<td>3</td>
<td>0.250</td>
<td>0.570</td>
<td>0.526</td>
</tr>
<tr>
<td>4, 5, 6</td>
<td>0.000</td>
<td>0.415</td>
<td>0.409</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subject</th>
<th>90% Unconditional BCN CI</th>
<th>90% Conditional BCN CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower</td>
<td>Upper</td>
</tr>
<tr>
<td>1, 2</td>
<td>0.108</td>
<td>0.690</td>
</tr>
<tr>
<td>3</td>
<td>0.044</td>
<td>0.570</td>
</tr>
<tr>
<td>4, 5, 6</td>
<td>0.006</td>
<td>0.415</td>
</tr>
</tbody>
</table>

Note: Hyperparameter estimate (MLE): \( \hat{b} = 3.9296 \).

### TABLE 3. Pump Failure Data

<table>
<thead>
<tr>
<th>System</th>
<th>X</th>
<th>t</th>
<th>r = \hat{\theta}_{Chown}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>94.320</td>
<td>.053</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>15.720</td>
<td>.064</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>62.889</td>
<td>.080</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
<td>125.760</td>
<td>.111</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>5.240</td>
<td>.573</td>
</tr>
<tr>
<td>6</td>
<td>19</td>
<td>31.440</td>
<td>.604</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1.048</td>
<td>.954</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1.048</td>
<td>.954</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>2.096</td>
<td>1.910</td>
</tr>
<tr>
<td>10</td>
<td>22</td>
<td>10.480</td>
<td>2.099</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Observation:</th>
<th>Pt Estimate</th>
<th>90% Confidence Interval</th>
<th>Interval Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classical</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classical</td>
<td>.053</td>
<td>.000</td>
<td>3.098</td>
</tr>
<tr>
<td>&quot;Naive G &amp; O</td>
<td>.056</td>
<td>.027</td>
<td>.114</td>
</tr>
<tr>
<td>&quot;Corrected&quot; G &amp; O</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Naive EB</td>
<td>.061</td>
<td>.026</td>
<td>.107</td>
</tr>
<tr>
<td>Unconditional BCN</td>
<td>-</td>
<td>.026</td>
<td>.109</td>
</tr>
<tr>
<td>Conditional BCN</td>
<td>-</td>
<td>(.0472)</td>
<td>(.9533)</td>
</tr>
<tr>
<td>Classical</td>
<td>.573</td>
<td>.004</td>
<td>4.032</td>
</tr>
<tr>
<td>&quot;Naive G &amp; O</td>
<td>.512</td>
<td>.207</td>
<td>1.258</td>
</tr>
<tr>
<td>&quot;Corrected&quot; G &amp; O</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Naive EB</td>
<td>.588</td>
<td>.195</td>
<td>1.154</td>
</tr>
<tr>
<td>Unconditional BCN</td>
<td>-</td>
<td>.158</td>
<td>(.0268)</td>
</tr>
<tr>
<td>Conditional BCN</td>
<td>-</td>
<td>.163</td>
<td>(.0298)</td>
</tr>
<tr>
<td>Classical</td>
<td>.604</td>
<td>.006</td>
<td>4.087</td>
</tr>
<tr>
<td>&quot;Naive G &amp; O</td>
<td>.577</td>
<td>.395</td>
<td>.861</td>
</tr>
<tr>
<td>&quot;Corrected&quot; G &amp; O</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Naive EB</td>
<td>.606</td>
<td>.401</td>
<td>.846</td>
</tr>
<tr>
<td>Unconditional BCN</td>
<td>-</td>
<td>.394</td>
<td>(.0437)</td>
</tr>
<tr>
<td>Conditional BCN</td>
<td>-</td>
<td>.398</td>
<td>(.0471)</td>
</tr>
<tr>
<td>Classical</td>
<td>2.099</td>
<td>.396</td>
<td>6.444</td>
</tr>
<tr>
<td>&quot;Naive G &amp; O</td>
<td>1.896</td>
<td>1.343</td>
<td>2.691</td>
</tr>
<tr>
<td>&quot;Corrected&quot; G &amp; O</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Naive EB</td>
<td>1.944</td>
<td>1.327</td>
<td>2.658</td>
</tr>
<tr>
<td>Unconditional BCN</td>
<td>-</td>
<td>1.230</td>
<td>(.0250)</td>
</tr>
<tr>
<td>Conditional BCN</td>
<td>-</td>
<td>1.311</td>
<td>(.0452)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A Sample Reuse Method For Accurate Parametric Empirical Bayes Confidence Intervals

Bradley P. Carlin and Alan E. Gelfand

Department of Statistics
Stanford University
Stanford, CA 94305

Office of Naval Research
Statistics & Probability Program Code 1111

November 7, 1989

Confidence interval; parametric empirical Bayes; parametric bootstrap; bias correction; conditional calibration.

PLEASE SEE FOLLOWING PAGE.
20. ABSTRACT

Parametric empirical Bayes methods of point estimation for a vector of unknown parameters date to the landmark paper of James and Stein (1961). The usual approach is to use the mean of the estimated posterior distribution of each parameter, where the estimation of the prior parameters ("hyperparameters") is accomplished through the marginal distribution of the data. While point estimates computed this way usually perform well, interval estimates based on the estimated posterior (called "naive" EB intervals) are not. They fail to account for the variability in the estimation of the hyperparameters, generally resulting in sub-nominal coverage probability in the "EB" sense defined in Morris (1983a).

In this paper we extend the work of Carlin and Gelfand (1989), who proposed a conditional bias correction method for developing EB intervals which corrects the deficiencies in the naive intervals. We show how bias correction can be implemented in general via a Type III parametric bootstrap procedure, a sample reuse method first employed by Laird and Louis (1987). Theoretical and simulation results indicate that intervals which are accurate with respect to nominal coverage ensue. We give two specific applications (to binomial test data and Poisson failure rate data) where we compute simultaneous point and bias corrected interval estimates.