ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2015)
77, *Part* 3, *pp.* 617–646

Frequentist accuracy of Bayesian estimates

Bradley Efron

Stanford University, USA

[Received October 2013. Final revision May 2014]

Summary. In the absence of relevant prior experience, popular Bayesian estimation techniques usually begin with some form of 'uninformative' prior distribution intended to have minimal inferential influence. The Bayes rule will still produce nice looking estimates and credible intervals, but these lack the logical force that is attached to experience-based priors and require further justification. The paper concerns the frequentist assessment of Bayes estimates. A simple formula is shown to give the frequentist standard deviation of a Bayesian point estimate. The same simulations as required for the point estimate also produce the standard deviation. Exponential family models make the calculations particularly simple and bring in a connection to the parametric bootstrap.

Keywords: Approximate bootstrap confidence intervals; General accuracy formula; Hierarchical and empirical Bayes; Markov chain Monte Carlo methods; Parametric bootstrap

1. Introduction

The past two decades have witnessed a greatly increased use of Bayesian techniques in statistical applications. Objective Bayes methods, based on neutral or uniformative priors of the type pioneered by Jeffreys, dominate these applications, carried forward on a wave of popularity for Markov chain Monte Carlo (MCMC) algorithms. Good references include Ghosh (2011), Berger (2006) and Kass and Wasserman (1996).

Suppose then that, having observed data x from a known parametric family $f_\mu(x)$, I wish to estimate $t(\mu)$, a parameter of particular interest. In the absence of relevant prior experience, I assign an uninformative prior $\pi(\mu)$, perhaps from the Jeffreys school. Applying Bayes rule yields $\hat{\theta}$, the posterior expectation of $t(\mu)$ given x:

$$\hat{\theta} = E\{t(\mu)|x\}. \tag{1.1}$$

How accurate is $\hat{\theta}$? The obvious answer, and the one that is almost always employed, is to infer the accuracy of $\hat{\theta}$ according to the Bayes posterior distribution of $t(\mu)$ given x. This would obviously be correct if $\pi(\mu)$ were based on genuine past experience. It is *not* so obvious for uninformative priors. I might very well like $\hat{\theta}$ as a point estimate, based on considerations of convenience, coherence, smoothness, admissibility or aesthetic Bayesian preference, but not trust what is after all a self-selected choice of prior as determining $\hat{\theta}$'s accuracy. Berger (2006) made this point at the beginning of his section 4.

As an alternative, this paper proposes computing the frequentist accuracy of $\hat{\theta}$, i.e. regardless of its Bayesian provenance, we consider $\hat{\theta}$ simply as a function of the data x and compute its frequentist variability.

Address for correspondence: Bradley Efron, Department of Statistics, Stanford University, Sequoia Hall, 390 Serra Mall, Stanford, CA 94305-4065, USA.
E-mail: brad@stat.stanford.edu

 1369–7412/15/77617

Our main result, which is presented in Section 2, is a general accuracy formula for the delta method standard deviation of $\hat{\theta}$: general in the sense that it applies to all prior distributions, uninformative or not. Even in complicated situations the formula is computationally inexpensive: the same MCMC calculations that give the Bayes estimate $\hat{\theta}$ also provide its frequentist standard deviation. A lasso-type example is used for illustration. Many of the examples that follow use Jeffreys priors; this is only for simplified exposition and is not a limitation of the theory.

In fact several of our examples will demonstrate near equality between Bayesian and frequentist standard deviations. That does not have to be so: remark 1 in Section 6 discusses a class of reasonable examples where the frequentist accuracy can be less than half of its Bayesian counterpart. Other examples will calculate frequentist standard deviations for situations where there is no obvious Bayesian counterpart, e.g. for the upper end point of a 95% credible interval.

The general accuracy formula takes on a particularly simple form when $f_\mu(x)$ represents a p-parameter exponential family: Section 3. Exponential family structure also allows us to substitute parametric bootstrap sampling for MCMC calculations, at least for uninformative priors. This has computational advantages. More importantly, it helps to connect Bayesian inference with the seemingly superfrequentist bootstrap world, which is a central theme of this paper.

The general accuracy formula provides frequentist standard deviations for Bayes estimators, but nothing more. Better inferences, in the form of second-order-accurate confidence intervals, are developed in Section 4, again in an exponential family bootstrap context. Section 5 uses the accuracy formula to compare hierarchical and empirical Bayes methods. The paper concludes with remarks, details and extensions in Section 6.

The frequentist properties of Bayes estimates is a venerable topic, that has been nicely reviewed in chapter 4 of Carlin and Louis (2000). Particular attention focuses on large sample behaviour, where 'the data swamp the prior' and $\hat{\theta}$ converges to the maximum likelihood estimator (see result 8 in section 4.7 of Berger (1985)), in which case the Bayes and frequentist standard deviations are nearly the same. Our accuracy formula provides some information about what happens *before* the data swamp the prior, though the present paper offers no proof of its superiority to standard asymptotic methods.

Some other important Bayesian-*cum*-frequentist topics are posterior and preposterior model checking as in Little (2006) or chapter 6 of Gelman *et al.* (1995), Bayesian consistency (Diaconis and Freedman, 1986), confidence matching priors, going back to Welch and Peers (1963), and empirical Bayes analysis as in Morris (1983). Johnstone and Silverman (2004) have provided, among much else, asymptotic bounds for the frequentist accuracy of empirical Bayes estimates.

Sensitivity analysis—modifying the prior as a check on the stability of posterior inference—is a staple of Bayesian model selection. The methods of this paper amount to modifying the *data* as a posterior stability check (see lemma 1 of Section 2). The implied suggestion here is to consider both techniques when the prior is in doubt.

The data sets and function `freqacc` are available from `http://statweb.stanford.edu/~brad/papers/jrss`.

2. General accuracy formula

We wish to estimate the frequentist accuracy of a Bayes posterior expectation $\hat{\theta} = E\{t(\mu)|x\}$ (1.1), where $t(\mu)$ is a parameter of particular interest. Here μ is an unknown parameter vector existing in parameter space Ω with prior density $\pi(\mu)$, whereas x is a sufficient statistic taking its values in, say, p-dimensional space,

$$x \in \mathcal{R}^p, \tag{2.1}$$

drawn from density $f_\mu(x)$ in a known parametric family

$$\mathcal{F} = \{f_\mu(x), \mu \in \Omega\}. \tag{2.2}$$

We write the expectation and covariance of x given μ as

$$x \sim (m_\mu, V_\mu) \tag{2.3}$$

with V_μ a $p \times p$ matrix. Denote the gradient of $\log\{f_\mu(x)\}$ with respect to x by

$$\alpha_x(\mu) = \nabla_x \log\{f_\mu(x)\} = \left(\cdots \frac{\partial}{\partial x_i} \log\{f_\mu(x)\} \cdots \right)^{\mathrm{T}}. \tag{2.4}$$

Lemma 1. The gradient of $\hat{\theta} = E\{t(\mu)|x\}$ with respect to x is the posterior covariance of $t(\mu)$ with $\alpha_x(\mu)$,

$$\nabla_x \hat{\theta} = \mathrm{cov}\{t(\mu), \alpha_x(\mu)|x\}. \tag{2.5}$$

Proof. Write $\hat{\theta} = A(x)/B(x)$ where

$$\begin{aligned} A(x) &= \int_\Omega t(\mu)\, \pi(\mu)\, f_\mu(x)\, \mathrm{d}\mu, \\ B(x) &= \int_\Omega \pi(\mu)\, f_\mu(x)\, \mathrm{d}\mu. \end{aligned} \tag{2.6}$$

Denoting the gradient operator ∇_x by primes, so $\alpha_x(\mu) = \log\{f_\mu(x)\}'$ (2.4), we calculate

$$\begin{aligned} A'(x) &= \int_\Omega t(\mu)\, \alpha_x(\mu)\, \pi(\mu)\, f_\mu(x)\, \mathrm{d}\mu, \\ B'(x) &= \int_\Omega \alpha_x(\mu)\, \pi(\mu)\, f_\mu(x)\, \mathrm{d}\mu. \end{aligned} \tag{2.7}$$

Using $(A/B)' = (A/B)(A'/A - B'/B)$ gives

$$\begin{aligned} \hat{\theta}' &= \hat{\theta} \left\{ \frac{\int_\Omega t(\mu)\, \alpha_x(\mu)\, \pi(\mu)\, f_\mu(x)\, \mathrm{d}\mu}{\int_\Omega t(\mu)\, \pi(\mu)\, f_\mu(x)\, \mathrm{d}\mu} - \frac{\int_\Omega \alpha_x(\mu)\, \pi(\mu)\, f_\mu(x)\, \mathrm{d}\mu}{\int_\Omega \pi(\mu)\, f_\mu(x)\, \mathrm{d}\mu} \right\} \\ &= \hat{\theta} \left[\frac{E\{t(\mu)\, \alpha_x(\mu)|x\}}{E\{t(\mu)|x\}} - E\{\alpha_x(\mu)|x\} \right] \\ &= E\{t(\mu)\, \alpha_x(\mu)|x\} - E\{t(\mu)|x\} E\{\alpha_x(\mu)|x\} \\ &= \mathrm{cov}\{t(\mu), \alpha_x(\mu)|x\}. \end{aligned}$$

□

A sufficient condition for the interchange of integration and differentiation in expression (2.7) is that $t(\mu)\, \pi(\mu)\, f'_\mu(x)$ be bounded in absolute value by a function $g(\mu, \tilde{x})$ having $\int_\Omega g(\mu, \tilde{x})\, \mathrm{d}\mu < \infty$ for $\tilde{x}$ in an open neighbourhood of x, and similarly for $\pi(\mu)\, f'_\mu(x)$. See section 2.4 of Casella and Berger (2002). Remark 2 of Section 6 presents a more computational derivation of lemma 1, where the crucial condition is only that the gradient $f'_\mu(\tilde{x})$ exists continuously in a neighbourhood of x.

Lemma 1 leads immediately to the *general accuracy formula*: general in the sense of applying to all choices of prior, not necessarily uninformative ones.

Theorem 1. The delta method approximation for the frequentist standard deviation of $\hat{\theta} = E\{t(\mu)|x\}$ is

$$\widehat{\mathrm{sd}} = [\mathrm{cov}\{t(\mu), \alpha_x(\mu)|x\}^{\mathrm{T}} V_{\hat{\mu}} \mathrm{cov}\{t(\mu), \alpha_x(\mu)|x\}]^{1/2}, \tag{2.8}$$

where $\hat{\mu}$ is the value of μ having $m_{\hat{\mu}} = x$.

Proof. Theorem 1 is an immediate consequence of lemma 1 and the usual delta method estimate of a statistic $s(x)$, as described for instance in section 4.6 of Rice (2007). Suppose for the sake of convenient notation that x is unbiased for μ, so $m_\mu = \mu$ in expression (2.3), and $\hat{\mu} = x$ in equation (2.8). Assuming that the gradient $s'(x) = \nabla_x(s)$ exists continuously in a neighbourhood of μ, a Taylor series expansion gives

$$s(x) = s(\mu) + s'(\mu)(x - \mu) + o(x - \mu). \tag{2.9}$$

The delta method (or 'propagation of errors' as it is known in the physical science literature) ignores the $o(x - \mu)$ term in equation (2.9) and approximates the standard deviation of $s(x)$ by

$$\mathrm{sd}_\mu(s) \doteq \{s'(\mu)^{\mathrm{T}} V_\mu s'(\mu)\}^{1/2}, \tag{2.10}$$

at the final step, plugging in an unbiased or maximum likelihood estimate (MLE) $\hat{\mu}$ for μ. Theorem 1 applies approximation (2.10) to $s(x) = \hat{\theta} = E\{t(\mu)|x\}$, using $s'(x) = \mathrm{cov}\{t(\mu), \alpha_x(\mu)|x\}$ (2.5). □

The delta method can be used to estimate bias as well as standard deviation, by extending equation (2.9) to a second-order Taylor series. Instead, the exponential family development of Section 4 provides second-order-accurate frequentist confidence intervals for $\hat{\theta}$, correcting for bias as well as other effects.

A useful special case of theorem 1 appears in Meneses *et al.* (1990). Fraser (1990), section 2, made use of $\nabla_x \log\{f_\mu(x)\}$ in likelihood-based procedures for calculating tail probabilities. The goal of Fraser (1990) is related to ours in the sense that likelihood methods enjoy a flat prior Bayesian interpretation. Fraser's work can be thought of as a continuation of the *matching priors* theory of Welch and Peers (1963), in which prior distributions are constructed to have favourable frequentist properties (as opposed to finding the frequentist properties of arbitrary priors, which is our goal here).

Several points about the general accuracy formula (2.6) are worth emphasizing.

(a) *Implementation*: suppose that

$$\{\mu_1, \mu_2, \mu_3, \ldots, \mu_B\} \tag{2.11}$$

is a sample of size B from the posterior distribution of μ given x. Each μ_i gives corresponding values of $t(\mu)$ and $\alpha_x(\mu)$ (2.4),

$$\begin{aligned} t_i &= t(\mu_i), \\ \alpha_i &= \alpha_x(\mu_i). \end{aligned} \tag{2.12}$$

Then $\bar{t} = \Sigma t_i / B$ approximates the posterior expectation $\hat{\theta}$, whereas

$$\widehat{\mathrm{cov}} = \sum_{i=1}^{B} (\alpha_i - \bar{\alpha})(t_i - \bar{t})/B \qquad \bar{\alpha} = \textstyle\sum \alpha_i / B \tag{2.13}$$

estimates the posterior covariance (2.5), so the same simulations that give $\hat{\theta}$ also provide

its frequentist standard deviation. (This assumes that $V_{\hat{\mu}}$ is easily available, as it will be in our applications.)

(b) *Posterior sampling*: the posterior sample $\{\mu_1, \mu_2, \ldots, \mu_B\}$ will typically be obtained via MCMC sampling, after a suitable burn-in period. The non-independence of the μs does not invalidate expression (2.13) but suggests that large values of B may be required for computational accuracy. The bootstrap-based posterior sampling method of Section 3 produces independent values μ_i. Independence permits simple assessments of the required size B; see expression (3.12) there.

(c) *Exponential families*: Section 3 shows that $\alpha_x(\mu)$ (2.4) has a simple form, not depending on x, in exponential families.

(d) *Factorization*: if

$$f_\mu(x) = g_\mu(x)\, h(x) \tag{2.14}$$

then the gradient

$$\nabla_x \log\{f_\mu(x)\} = \nabla_x \log\{g_\mu(x)\} + \nabla_x \log\{h(x)\}. \tag{2.15}$$

The last term does not depend on μ, so

$$\operatorname{cov}[t(\mu), \nabla_x \log\{f_\mu(x)\}|x] = \operatorname{cov}[t(\mu), \nabla_x \log\{g_\mu(x)\}|x]$$

and we can take

$$\alpha_x(\mu) = \nabla_x \log\{g_\mu(x)\} \tag{2.16}$$

in lemma 1 and theorem 1.

(e) *Sufficiency*: if $x = (y, z)$ where x is p dimensional and $y = Y(x)$ is a q-dimensional sufficient statistic, we can write $f_\mu(x) = g_\mu(y)\, h(z)$ and

$$\alpha_x(\mu) = \nabla_x \log\{f_\mu(x)\} = \nabla_x \log\{g_\mu(y)\} + \nabla_x \log\{h(z)\}. \tag{2.17}$$

As in equation (2.15), the last term does not depend on μ so we can take $\alpha_x(\mu) = \nabla_x \log\{g_\mu(y)\}$. Letting $\alpha_y(\mu) = \nabla_y \log\{g_\mu(y)\}$, a q-dimensional vector,

$$\alpha_x(\mu) = {Y'}^{\mathrm{T}}\, \alpha_y(\mu), \tag{2.18}$$

where Y' is the $q \times p$ matrix $(\partial y_i / \partial x_j)$. From equation (2.8) we obtain

$$\widehat{\mathrm{sd}} = [\operatorname{cov}\{t(\mu), \alpha_y(\mu)|y\}^{\mathrm{T}} Y' V_{\hat{\mu}} {Y'}^{\mathrm{T}} \operatorname{cov}\{t(\mu), \alpha_y(\mu)|y\}]^{1/2}. \tag{2.19}$$

Note that $Y' V_{\hat{\mu}} {Y'}^{\mathrm{T}}$ is the delta method estimate of the covariance matrix of y when μ equals $\hat{\mu}$. In this approximate sense theorem 1 automatically accounts for sufficiency. However, we can avoid the approximation if in the first place we work with y and its actual covariance matrix. (This will be so in the exponential family set-up of Section 3.) More importantly, working with y makes $\widehat{\mathrm{cov}}$ in expression (2.13) lower dimensional and yields better estimation properties when substituted into equation (2.8).

(f) *Vector parameter of interest*: lemma 1 and theorem 1 apply also to the case where the target parameter $t(\mu)$ is vector valued, say K dimensional, as is $\hat{\theta} = E\{t(\mu)|x\}$. Then $\nabla_x \hat{\theta}$ and $\operatorname{cov}\{t(\mu), \alpha_x(\mu)|x\}$ in equation (2.5) become $p \times K$ matrices, yielding $K \times K$ approximate frequentist covariance matrix $\widehat{\mathrm{var}}$ for $\hat{\theta} = E\{t(\mu)|x\}$,

$$\widehat{\mathrm{var}} = \operatorname{cov}\{t(\mu), \alpha_x(\mu)|x\}^{\mathrm{T}} V_{\hat{\mu}} \operatorname{cov}\{t(\mu), \alpha_x(\mu)|x\}, \tag{2.20}$$

with $\alpha_x(\mu)$ and $V_{\hat{\mu}}$ the same as before.

(g) *Discrete statistic* x: suppose that $\mathcal{F}$ in expression (2.2) is the one-dimensional Poisson family $f_\mu(x) = \exp(-\mu)\,\mu^x/x!$, with x a non-negative integer. We can still calculate $\alpha_x(\mu) = \log(\mu)$ (2.4) (ignoring the term due to $x!$, as in equation (2.15)). For μ greater than, say, 10, the Poisson distribution ranges sufficiently widely to smooth over its discrete nature, and we can expect formula (2.8) to apply reasonably well. Section 5 discusses a multi-dimensional discrete application.

(h) *Standard deviation bias correction*: replacing $\mathrm{cov}\{t(\mu), \alpha_x(\mu)|x\}$ in equation (2.8) with its nearly unbiased estimate $\widehat{\mathrm{cov}}$ (2.13) upwardly biases the standard deviation estimate. Remark 4 of Section 6 discusses a simple bias correction. Bias was negligible in the numerical examples that follow.

As an example of theorem 1 in action, we shall consider the *diabetes data* of Efron *et al.* (2004): $n = 442$ diabetes patients each have had observed a vector $\mathbf{x}$ of $p = 10$ predictor variables (age, sex, body mass index, blood pressure and six blood serum measurements),

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{i10}) \qquad \text{for } i = 1, 2, \ldots, n = 442, \tag{2.21}$$

and also a response variable y_i measuring disease progression at 1 year after entry. Standardizing the predictors and response variables suggests a normal linear model

$$\mathbf{y} = \mathbf{X}\alpha + \mathbf{e} \qquad \mathbf{e} \sim \mathcal{N}_n(0, \mathbf{I}). \tag{2.22}$$

Here $\mathbf{X}$ is the $n \times p$ matrix having ith row $\mathbf{x}_i$, whereas $\mathbf{y}$ is the vector of n responses.

Park and Casella (2008) considered applying a Bayesian version of the lasso (Tibshirani, 1996) to the diabetes data. In terms of our model (2.22) (they did not standardize the response) Park and Casella took the prior distribution for α to be

$$\pi(\alpha) = \exp\{-\lambda L_1(\alpha)\}, \tag{2.23}$$

with $L_1(\alpha)$ the L_1-norm $\Sigma_1^{10}|\alpha_j|$, and λ having value (in our standardized set-up)

$$\lambda = 0.37. \tag{2.24}$$

The Laplace-type prior (2.23) results in the posterior mode of α given $\mathbf{y}$ coinciding with the lasso estimate

$$\hat{\alpha}_\lambda = \arg\min_\alpha \{\|\mathbf{y} - \mathbf{X}\alpha\|^2/2 + \lambda L_1(\alpha)\}, \tag{2.25}$$

as pointed out in Tibshirani (1996). The choice $\lambda = 0.37$ was obtained from marginal maximum likelihood considerations. In this sense Park and Casella's analysis is *empirical* Bayesian, but we shall ignore that here and assume prior (2.23)–(2.24) is preselected. (The lasso itself plays no role in their calculations or those here except as motivation.)

An MCMC algorithm was used to produce (after burn-in) $B = 10\,000$ samples α_i from the posterior distribution $\pi(\alpha|\mathbf{y})$, under assumptions (2.22)–(2.24):

$$\{\alpha_i, i = 1, 2, \ldots, B = 10\,000\}. \tag{2.26}$$

From these we can approximate the Bayes posterior expectation $\hat{\theta} = E(\gamma|\mathbf{y})$ for any parameter of interest $\gamma = t(\alpha)$,

$$\hat{\theta} = \sum_{i=1}^{B} t(\alpha_i)/B. \tag{2.27}$$

Note that it is helpful here and in what follows to denote the parameter of interest as $\gamma = t(\mu)$ with $\hat{\theta} = E(\gamma|x)$ indicating its posterior expectation.

We can now apply theorem 1 to estimate the frequentist standard deviation of $\hat{\theta}$. In terms of the general notation (2.2), μ becomes α, and we can take x to be the sufficient statistic $\hat{\beta} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$ in model (2.22). (We could take $x = \hat{\alpha} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$, but these choices make α the natural parameter vector and $\hat{\beta}$ the sufficient statistic in the exponential family form (3.1).) Section 3 shows that $\alpha_x(\mu)$ (2.4) equals α in an exponential family. With $\widehat{\mathrm{cov}}$ computed as in expression (2.13), the computational form of theorem 1 yields frequentist standard deviation

$$\widehat{\mathrm{sd}} = (\widehat{\mathrm{cov}}^{\mathrm{T}} G \widehat{\mathrm{cov}})^{1/2} \qquad (G = \mathbf{X}^{\mathrm{T}}\mathbf{X}), \tag{2.28}$$

since G is the variance matrix V of $\hat{\beta}$.

As a univariate 'parameter of special interest', consider estimating

$$\gamma_{125} = \mathbf{x}_{125}\alpha, \tag{2.29}$$

the diabetes progression for patient 125. (Patient 125 fell near the centre of the y response scale.) The 10 000 values $\hat{\gamma}_{125,i} = \mathbf{x}_{125}\alpha_i$ were nearly normally distributed:

$$\mathcal{N}(0.248, 0.072^2). \tag{2.30}$$

Formula (2.28) gave frequentist standard deviation 0.071 for the posterior expectation of γ_{125}, $\hat{\theta}_{125} = 0.248 = \Sigma\hat{\gamma}_{125,i}/10\,000$, which is almost the same as the posterior standard deviation, but having quite a different interpretation. The near equality here is no fluke but can turn out differently for other linear combinations $\gamma = \mathbf{x}\alpha$; see remark 1 of Section 6.

Suppose we are interested in the posterior cumulative distribution function (CDF) of γ_{125}. For a given value c define

$$t_c(\alpha) = \begin{cases} 1 & \text{if } \mathbf{x}_{125}\alpha \leqslant c, \\ 0 & \text{if } \mathbf{x}_{125}\alpha > c \end{cases} \tag{2.31}$$

so $E\{t_c(\alpha)|\mathbf{y}\} = \Pr(\gamma_{125} \leqslant c|\mathbf{y})$. The MCMC sample (2.26) provides $B = 10\,000$ posterior values t_{ci}, from which we obtain the estimated CDF(c) value $\Sigma_1^B t_{ci}/B$ and its standard deviation (2.8); for example $c = 0.3$ gives

$$\Pr(\gamma_{125} \leqslant 0.3|\mathbf{y}) = 0.762 \pm 0.304, \tag{2.32}$$

0.304 being the frequentist standard deviation of the posterior Bayes CDF 0.762.

The bold curve in Fig. 1 traces the posterior CDF of γ_{125}. The broken vertical bars indicate ± 1 frequentist standard deviation. If we take prior (2.23) literally then the CDF curve is exact but, if not, the large frequentist standard errors suggest cautious interpretation; in the same way we might react to a disturbing sensitivity analysis on the choice of prior.

The CDF curve equals 0.90 at $\hat{c} = 0.342$, this being the upper end point of a one-sided Bayes 90% credible interval. The frequentist standard deviation of $\hat{c}$ is 0.069 (obtained from $\widehat{\mathrm{sd}}\{\mathrm{CDF}(\hat{c})\}$ divided by the posterior density at $\hat{c}$, the usual delta method approximation), giving coefficient of variation $0.069/0.342 = 0.20$.

For θ_{125} itself we could compare the frequentist standard deviation 0.071 with its Bayes posterior counterpart 0.072 (2.30). No such comparison is possible for the posterior CDF estimates: the CDF curve in Fig. 1 is exact under prior (2.23)–(2.24). We might add a hierarchical layer of Bayesian assumptions in front of prior (2.23)–(2.24) in order to assess the curve's

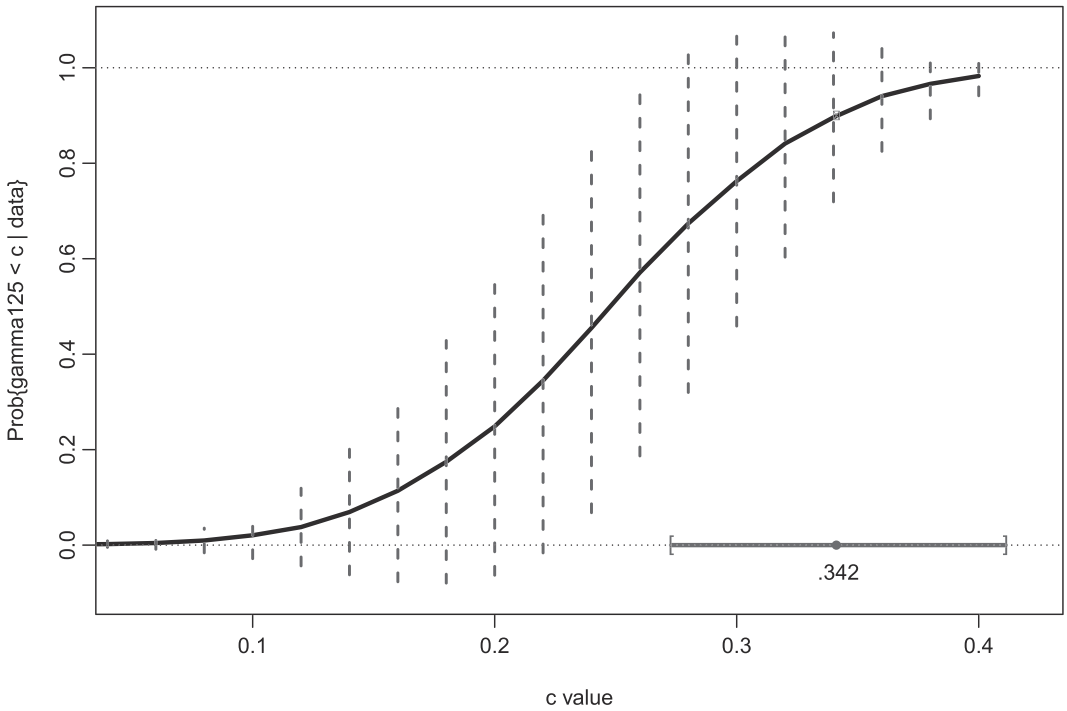

Fig. 1. Posterior CDF of γ_{125} (2.29) (———) for the diabetes data (¦, ±1 frequentist standard error): the estimated curve is quite uncertain from a frequentist viewpoint; the upper 0.90 value $\hat{c} = 0.342$ has frequentist standard error 0.069, as indicated by the horizontal bar

variability, but it is not obvious how to do so here. (Park and Casella (2008), section 3.2, made one suggestion.)

The frequentist error bars of Fig. 1 extend below 0 and above 1, a reminder that standard deviations are a relatively crude inferential tool. Section 4 discusses more sophisticated frequentist methods.

3. A bootstrap version of the general formula

A possible disadvantage of Section 2's methodology is the requirement of a posterior sample $\{\mu_1, \mu_2, \ldots, \mu_B\}$ from $\pi(\mu|x)$ (2.11). This section discusses a parametric bootstrap approach to the general accuracy formula that eliminates posterior sampling, at the price of less generality: a reduction of scope to exponential families and to priors $\pi(\mu)$ that are at least roughly uninformative. On the other hand, the bootstrap methodology makes the computational error analysis, i.e. the choice of the number of replications B, straightforward and, more importantly, helps to connect Bayesian and frequentist points of view.

A p-parameter exponential family $\mathcal{F}$ can be written as

$$\mathcal{F}: \{f_\alpha(\hat{\beta}) = \exp\{\alpha^{\mathrm{T}}\hat{\beta} - \psi(\alpha)\}\, f_0(\hat{\beta}), \alpha \in \mathcal{A}\}. \tag{3.1}$$

Here α is the natural or canonical parameter vector, and $\hat{\beta}$ is the p-dimensional sufficient statistic. The *expectation parameter* $\beta = E_\alpha(\hat{\beta})$ is a one-to-one function of α, say $\beta = A(\alpha)$, with $\hat{\beta}$ equalling the MLE of β. The parameter space $\mathcal{A}$ for α is a subset of $\mathcal{R}^p$, p-dimensional space,

as is the corresponding space for β. The function $\psi(\alpha)$ provides the multiplier necessary for $f_\alpha(\hat\beta)$ to integrate to 1.

In terms of the generic notation (2.1)–(2.2), α is μ and $\hat\beta$ is x. The expectation and covariance of $\hat\beta$ given α,

$$\hat\beta \sim (\beta, V_\alpha), \tag{3.2}$$

can be obtained by differentiating $\psi(\alpha)$.

The general accuracy formula (2.8) takes a simplified form in exponential families.

Theorem 2. The delta method approximation for the frequentist standard deviation of $\hat\theta = E\{t(\alpha)|\hat\beta\}$ in exponential family (3.1) is

$$\widehat{\mathrm{sd}} = [\mathrm{cov}\{t(\alpha), \alpha|\hat\beta\}^{\mathrm{T}} V_{\hat\alpha}\, \mathrm{cov}\{t(\alpha), \alpha|\hat\beta\}]^{1/2}, \tag{3.3}$$

where $\hat\alpha$, the natural parameter vector corresponding to $\hat\beta$, is the MLE of α.

Proof. The gradient $\nabla_x \log\{f_\mu(x)\}$ in expression (2.4) is now

$$\begin{aligned}\nabla_{\hat\beta} \log\{f_\alpha(\hat\beta)\} &= \nabla_{\hat\beta}[\alpha^{\mathrm{T}}\hat\beta - \psi(\alpha) + \log\{f_0(\hat\beta)\}] \\ &= \alpha + \nabla_{\hat\beta} \log\{f_0(\hat\beta)\}.\end{aligned} \tag{3.4}$$

The final term does not depend on α so, as in equation (2.15), what was called $\alpha_x(\mu)$ in expression (2.4) becomes simply α, reducing equation (2.8) to equation (3.3). □

Parametric bootstrap resampling can be employed to calculate both $\hat\theta$ and $\widehat{\mathrm{sd}}$, as suggested in Efron (2012). We independently resample B times from the member of $\mathcal{F}$ having parameter vector α equal to $\hat\alpha$:

$$f_{\hat\alpha}(\cdot) \to \{\beta_1, \beta_2, \ldots, \beta_i, \ldots, \beta_B\} \tag{3.5}$$

(β_i being short for the conventional bootstrap notation $\hat\beta_i^*$). Each β_i gives a corresponding natural parameter vector $\alpha_i = A^{-1}(\beta_i)$. Let $\pi_i = \pi(\alpha_i)$, and define the 'conversion factor'

$$R_i = f_{\alpha_i}(\hat\beta)/f_{\hat\alpha}(\beta_i), \tag{3.6}$$

the ratio of the likelihood to the bootstrap density. (See equations (3.13)–(3.15) for the evaluation of R_i.)

The discrete distribution that puts weight

$$p_i = \pi_i R_i \Big/ \sum_{j=1}^{B} \pi_j R_j \tag{3.7}$$

on α_i, for $i = 1, 2, \ldots, B$, approximates the conditional distribution of α given $\hat\beta$. To see this let $t_i = t(\alpha_i)$ and $\hat\theta_B = \Sigma_1^B t_i p_i$, so

$$\hat\theta_B = \frac{\sum_{i=1}^{B} t_i \pi_i f_{\alpha_i}(\hat\beta) \Big/ f_{\hat\alpha}(\beta_i)}{\sum_{i=1}^{B} \pi_i f_{\alpha_i}(\hat\beta) \Big/ f_{\hat\alpha}(\beta_i)}. \tag{3.8}$$

Since the β_i are drawn from bootstrap density $f_{\hat\alpha}(\cdot)$, equation (3.8) represents an importance sampling estimate of

$$\int_{\mathcal{A}} t(\alpha)\,\pi(\alpha)\,f_\alpha(\hat\beta)\;\mathrm{d}\alpha \Big/ \int_{\mathcal{A}} \pi(\alpha)\,f_\alpha(\hat\beta), \tag{3.9}$$

which equals $E\{t(\alpha)|\hat\beta\}$.

The same argument applies to any posterior calculation. In particular, $\mathrm{cov}\{t(\alpha),\alpha|\hat\beta\}$ in expression (3.3) is approximated by

$$\widehat{\mathrm{cov}} = \sum_{i=1}^{B} p_i(\alpha_i-\bar\alpha)(t_i-\hat\theta) \qquad \bar\alpha = \sum p_i\alpha_i,\ \hat\theta = \sum p_i t_i. \tag{3.10}$$

Implementing theorem 2 now follows three algorithmic steps.

Step 1: generate a parametric bootstrap sample $\beta_1,\beta_2,\ldots,\beta_B$ (3.5).
Step 2: for each β_i calculate α_i, $t_i = t(\alpha_i)$ and p_i (3.7).
Step 3: compute $\widehat{\mathrm{cov}}$ (3.10).

Then $\hat\theta_B = \Sigma\, p_i t_i$ approximates $\hat\theta = E\{t(\alpha)|\hat\beta\}$ and has delta method frequentist standard deviation

$$\widehat{\mathrm{sd}} = (\widehat{\mathrm{cov}}^{\mathrm{T}} V_{\hat\alpha}\,\widehat{\mathrm{cov}})^{1/2}. \tag{3.11}$$

(The matrix $V_{\hat\alpha}$ can be replaced by the empirical covariance matrix of $\beta_1,\beta_2,\ldots,\beta_B$ or, with one further approximation, by the inverse of the covariance matrix of $\alpha_1,\alpha_2,\ldots,\alpha_B$.) Remark 3 of Section 6 develops an alternative expression for $\widehat{\mathrm{sd}}$. In what follows, $\hat\theta_B$ is called simply $\hat\theta$.

An MCMC implementation sample $\{\mu_i,\ i=1,2,\ldots,B\}$ (2.11) approximates a multidimensional posterior distribution by an equally weighted distribution on B non-independent points. The bootstrap implementation (3.5)–(3.7) puts *unequal* weights on B independent and identically distributed (IID) points.

Independent resampling permits a simple analysis of 'internal accuracy', the error due to stopping at B resamples rather than letting $B\to\infty$. Define $P_i = \pi_i R_i$ and $Q_i = t_i P_i = t_i\pi_i R_i$. Since the pairs (P_i, Q_i) are independently resampled, standard delta method calculations show that $\hat\theta = \Sigma\, Q_i/\Sigma\, P_i$ has internal squared coefficient of variation approximately

$$\widehat{\mathrm{cv}}^2_{\mathrm{int}} = \frac{1}{B}\sum_{i=1}^{B}\left(\frac{Q_i}{\bar Q} - \frac{P_i}{\bar P}\right)^2 \Big/ B, \tag{3.12}$$

$\bar Q = \Sigma\, Q_i/B$ and $\bar P = \Sigma\, P_i/B$. See remark 3 of Appendix A.

There are two sources of approximation in applying the general accuracy formula: Monte Carlo error due to stopping at B replications, and delta method error in estimating the true standard deviation. For bootstrap sampling, formula (3.12) assesses the Monte Carlo error. The better bootstrap confidence intervals of Section 4 improve on the inferential approximations of the delta method.

The conversion factor R_i (3.6) can be defined for any family $\{f_\alpha(\hat\beta)\}$, but it has a simple expression in exponential families:

$$R_i = \xi(\alpha_i)\,\exp\{\Delta(\alpha_i)\}, \tag{3.13}$$

where $\Delta(\alpha)$ is the 'half-deviance difference'

$$\Delta(\alpha) = (\alpha-\hat\alpha)^{\mathrm{T}}(\beta+\hat\beta) - 2\{\psi(\alpha)-\psi(\hat\alpha)\}, \tag{3.14}$$

and, to a good approximation (Efron (2012), lemma 1),

Table 1. Cell infusion data†

Infusion proportion	*Results for the following numbers of days:*				
	1	*2*	*3*	*4*	*5*
1	5/31	3/28	20/45	24/47	29/35
2	15/77	36/78	43/71	56/71	66/74
3	48/126	68/116	145/171	98/119	114/129
4	29/92	35/52	57/85	38/50	72/77
5	11/53	20/52	20/48	40/55	52/61

†Human cell colonies were infused with mouse nuclei in five different proportions, over time periods varying from 1 to 5 days, and observed to see whether they did or did not thrive. The table displays the number thriving over the number of colonies. For example, five of the 31 colonies in the lowest infusion–days category thrived.

$$\xi(\alpha) = 1/\pi^{\text{Jeff}}(\alpha), \tag{3.15}$$

with $\pi^{\text{Jeff}}(\alpha) = |V_\alpha|^{1/2}$, Jeffreys invariant prior for α. If our prior $\pi(\alpha)$ *is* $\pi^{\text{Jeff}}(\alpha)$ then

$$\pi_i R_i = \exp\{\Delta(\alpha_i)\}. \tag{3.16}$$

The bootstrap distribution $f_{\hat\alpha}(\cdot)$ locates its resamples α_i near the MLE $\hat\alpha$. A working definition of an *informative prior* $\pi(\alpha)$ might be a prior that places substantial probability far from $\hat\alpha$. In that case, R_i is liable to take on enormous values, destabilizing the importance sampling calculations. Park and Casella's (2008) prior (2.23)–(2.24) for the diabetes data would be a poor choice for bootstrap implementation (though this difficulty can be mitigated by recentring the parametric bootstrap resampling distribution).

Table 1 displays the *cell infusion data*, which we shall use to illustrate bootstrap implementation of the general accuracy formula. Human muscle cell colonies were infused with mouse nuclei. Five increasing infusion proportions of mouse nuclei were tried, cultured over time periods ranging from 1 to 5 days, and observed to find whether they thrived or not. Table 1 shows that 52 of the 61 colonies in the highest proportion–days category thrived, etc.

Letting (s_{jk}, n_{jk}) be the number of successes and number of colonies in the jkth cell, we assume independent binomial variation:

$$s_{jk} \overset{\text{ind}}{\sim} \text{Bi}(n_{jk}, \xi_{jk}) \qquad j = 1, 2, \ldots, 5, \quad k = 1, 2, \ldots, 5. \tag{3.17}$$

An additive logistic regression model fitted the data reasonably well:

$$\text{logit}(\xi_{jk}) = \alpha_0 + \alpha_1 I_j + \alpha_2 I_j^2 + \alpha_3 D_k + \alpha_4 D_k^2, \tag{3.18}$$

with I_j the infusion proportions $1, 2, \ldots, 5$, and D_k the days $1, 2, \ldots, 5$. Model (3.18) is a five-parameter exponential family (3.1).

For our parameter of special interest $t(\alpha)$ we shall take

$$\gamma = \sum_{j=1}^{5} \xi_{j5} \Big/ \sum_{j=1}^{5} \xi_{j1}, \tag{3.19}$$

which is the ratio of overall probability of success on day 5 compared with day 1, and calculate its

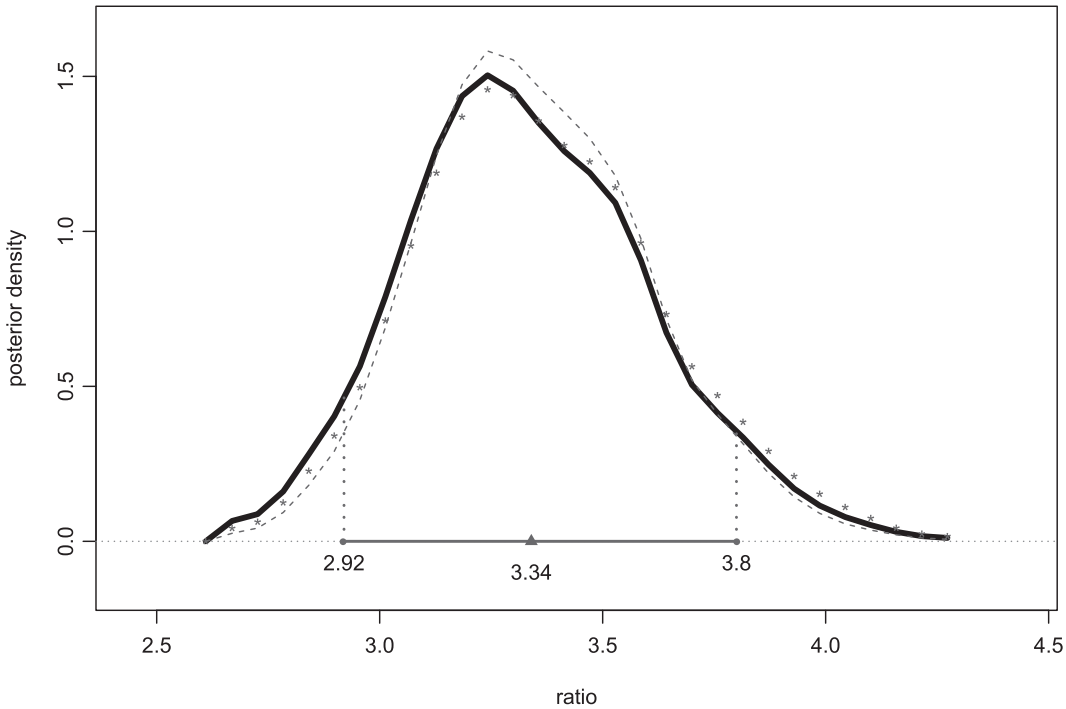

Fig. 2. Posterior density of ratio γ (3.19) given the cell infusion data, for the binomial model (3.17)–(3.18) and Jeffreys prior $\pi^{\text{Jeff}}(\alpha)$ (- - - - -, posterior density of γ by using the conjugate prior density (3.23) with $c_0 = 0.2$ and b_0 equal to the MLE $\hat{\beta}$; *, raw unweighted bootstrap density): from $B = 2000$ parametric bootstrap replications (3.20), posterior expectation 3.34 has frequentist sd = 0.273; the line segment shows central 0.90 credible interval [2.92, 3.80]; the frequentist standard deviation of 0.90 content is 0.042

posterior distribution assuming Jeffreys prior $\pi^{\text{Jeff}}(\alpha)$ on α. Warning: Jeffreys prior is convenient for illustrative purposes here but can be dangerous to use in multi-dimensional situations. An alternative analysis based on conjugate priors appears below.

$B = 2000$ parametric bootstrap samples were generated according to

$$s^*_{jk} \overset{\text{ind}}{\sim} \text{Bi}(n_{jk}, \hat{\xi}_{jk}), \qquad j = 1, 2, \ldots, 5, \quad k = 1, 2, \ldots, 5, \tag{3.20}$$

where $\hat{\xi}_{jk}$ is the MLE of ξ_{jk} obtained from model (3.18). These gave bootstrap MLEs $\alpha_1, \alpha_2, \ldots, \alpha_i, \ldots, \alpha_{2000}$ and corresponding bootstrap estimates $\gamma_i = t(\alpha_i)$ as in equation (3.19). The weights p_i (3.7) that convert the bootstrap sample into a posterior distribution are

$$p_i = \exp(\Delta_i) \Big/ \sum_{j=1}^{2000} \exp(\Delta_j) \tag{3.21}$$

according to equation (3.16), with Δ_i the half-binomial deviance difference (3.14); see remark 5 in Section 6.

The bold curve in Fig. 2 is the estimated posterior density, i.e. a smoothed version of the discrete distribution putting weight p_i on $\gamma_i = t(\alpha_i)$. Its expectation

$$\hat{\theta}_B = \sum_1^B p_i t(\alpha_i) = 3.34 \tag{3.22}$$

Table 2. Posterior expectation and standard deviation of γ (3.19) for a Jeffreys prior and six choices of c_0 for conjugate prior (3.23), $b_0 = \hat{\beta}$†

	Jeffreys prior	*Conjugate prior for the following values of* c_0*:*						*Bootstrap*
		0.005	*0.01*	*0.025*	*0.05*	*0.1*	*0.2*	
Expectation	3.335	3.348	3.348	3.349	3.349	3.349	3.350	3.361
Standard deviation	0.272	0.274	0.273	0.271	0.268	0.263	0.252	0.270

†At the right are the expectation and standard deviation for the unweighted bootstrap distribution.

is a Monte Carlo estimate of the posterior expectation of γ given the data. ($B = 2000$ resamples were excessive, formula (3.12) giving internal coefficient of variation only 0.002.)

How accurate is $\hat{\theta}$? Formula (3.11) yields $\widehat{\mathrm{sd}} = 0.273$ as its frequentist standard deviation. This is almost the same as the Bayes posterior standard deviation $\{\Sigma\, p_i(\gamma_i - \hat{\theta})^2\}^{1/2} = 0.272$.

In this case we can see why the Bayesian and frequentist standard deviations might be so similar: the Bayes posterior density is nearly the same as the raw bootstrap density (weight $1/B$ on each value γ_i). This happens whenever the parameter of interest has low correlation with the weights p_i (lemma 3 of Efron (2014)). The bootstrap estimate of standard deviation $\{\Sigma\,(\gamma_i - \bar{\gamma})^2\}^{1/2}$ equals 0.270, and it is not surprising that both the Bayes posterior standard deviation and the frequentist delta method standard deviation are close to 0.270.

Integrating the author's full curve in Fig. 2 gives [2.92, 3.80] as the 0.90 central credible interval for γ. Defining t_i to be 1 or 0 as γ_i does or does not fall into this interval, formula (3.11) yields $\widehat{\mathrm{sd}} = 0.042$ for the frequentist standard deviation of the interval's content. The two end points have standard deviations 0.22 and 0.31. More interestingly, their frequentist correlation (calculated by using equation (2.20); see remark 6 of Section 6) is 0.999. This strongly suggests that replications of the muscle data experiment would show the 0.90 credible interval shifting left or right, without much change in length.

As an alternative to $\pi^{\mathrm{Jeff}}(\alpha)$ we also considered conjugate priors for the exponential family (3.17)–(3.18). In terms of expression (3.1), conjugate priors have the form

$$\pi_{c_0,b_0}(\alpha) = \exp[c_0\{\alpha^{\mathrm{T}} b_0 - \psi(\alpha)\}] \tag{3.23}$$

(Diaconis and Ylvisaker, 1979). The $p \times p$ second-derivative matrix of $-\log\{\pi_{c_0,b_0}(\alpha)\}$ is $c_0\,\ddot{\psi}(\alpha)$, with $\ddot{\psi}(\alpha) = (\partial^2\psi/\partial\alpha_i\,\partial\alpha_j)$, compared with $\ddot{\psi}(\alpha)$ for $-\log\{f_\alpha(\hat{\beta})\}$, so small values of c_0 make $\pi_{c_0,b_0}(\alpha)$ more diffuse than the distribution of the MLE $\hat{\alpha}$. Spiegelhalter and Smith (1982), writing in a model selection context, recommended setting b_0 equal to $\hat{\beta}$, the MLE of $\beta = E_\alpha(\hat{\beta})$, with c_0 perhaps $1/n$ for an IID sample of size n. For the non-IID data of Table 1, rough information calculations suggest c_0 of the order of 0.01.

Posterior expectations and standard deviations (*not* frequentist standard deviations from the general accuracy formula) are given in Table 2, for six choices of c_0 ranging from 0.005 to 0.2. These do not differ much from each other or from the Jeffreys moments, and all are close to the unweighted bootstrap values.

(R function `freqacc` calculates frequentist standard deviations for Bayes estimates that are obtained either by MCMC sampling as in Section 2 or by bootstrap reweighting as here; the function assumes exponential family form (3.1).)

4. Improved inferences

The general accuracy formula of theorem 1 and theorem 2 computes frequentist standard deviations for Bayesian estimates. Standard deviations are a good start but not the last word in assessing the accuracy of a point estimator. A drawback is apparent in Fig. 1, where the standard error bars protrude beyond the feasible interval [0, 1].

This section concerns bootstrap methods that provide better frequentist inference for Bayesian estimates. A straightforward bootstrap approach would begin by obtaining a preliminary set of resamples, say

$$f_{\hat{\alpha}}(\cdot) \to \hat{b}_1^*, \hat{b}_2^*, \ldots, \hat{b}_K^* \tag{4.1}$$

in the exponential family set-up (3.1), for each $\hat{b}_k^*$ calculating $\hat{\theta}_k^* = \hat{E}\{t(\alpha)|\hat{b}_k^*\}$, the posterior expectation of $t(\alpha)$ given sufficient statistic $\hat{b}_k^*$, and finally using $\{\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_K^*\}$ to form a bootstrap confidence interval corresponding to the point estimate $\hat{\theta} = E\{t(\alpha)|\hat{\beta}\}$, perhaps the $\mathrm{BC_a}$ interval (Efron, 1987). By construction, such intervals would *not* protrude beyond [0, 1] in the equivalent of Fig. 1 and would take into account bias and interval asymmetry as well as standard deviation.

The roadblock to the straightforward approach is excessive computation. Bootstrap confidence intervals require K, the number of replicates, to be of the order of 1000. Each of these would require further simulations, $\{\mu_1, \mu_2, \ldots, \mu_B\}$ as in expression (2.11) or $\{\beta_1, \beta_2, \ldots, \beta_B\}$ as in expression (3.5), B also exceeding 1000, to calculate the $\hat{\theta}_k^*$ accurately. (The change in notation from expression (3.5) to expression (4.1) is intended to emphasize that each $\hat{b}_k^*$ needs to be followed, at least in the straightforward approach, by its own second-level bootstrap sample (3.5).)

A short-cut method for bootstrap confidence calculations that, like theorems 1 and 2, requires *no* additional replications will be developed next. The short cut requires exponential family structure (3.1) but otherwise applies equally to MCMC or bootstrap implementation (2.11) or (3.5).

The Bayes theorem says that the posterior density $g(\alpha|\hat{\beta})$ corresponding to exponential family density $f_\alpha(\hat{\beta})$ (3.1) is

$$g(\alpha|\hat{\beta}) = \pi(\alpha)\, f_\alpha(\hat{\beta}) / f(\hat{\beta}) \qquad f(\hat{\beta}) = \int_{\mathcal{A}} \pi(\alpha)\, f_\alpha(\hat{\beta})\, \mathrm{d}\alpha. \tag{4.2}$$

Suppose that now we change the observed sufficient statistic vector $\hat{\beta}$ to a different value b.

Lemma 2. The posterior distributions corresponding to exponential family $\mathcal{F}$ form an exponential family $\mathcal{G}$,

$$\mathcal{G} = \{g(\alpha|b) = \exp\{(b-\hat{\beta})^{\mathrm{T}}\alpha - \phi(b)\}\, g(\alpha|\hat{\beta}) \text{ for } b - \hat{\beta} \in \hat{\mathcal{B}}\}, \tag{4.3}$$

where

$$\exp\{\phi(b)\} = \int_{\mathcal{A}} \exp\{(b-\hat{\beta})^{\mathrm{T}}\alpha\}\, g(\alpha|\hat{\beta})\, \mathrm{d}\alpha. \tag{4.4}$$

$\mathcal{G}$ is a p-parameter exponential family with roles reversed from $\mathcal{F}$; now α is the sufficient statistic and b the natural parameter vector; $\hat{\mathcal{B}}$ is the convex set of vectors $b - \hat{\beta}$ for which the integral in equation (4.4) is finite.

($\mathcal{G}$ is not the familiar *conjugate family* (Diaconis and Ylvisaker, 1979), though there are connections.)

Proof. From expression (3.1) we compute

$$\begin{aligned} g(\alpha|b) &= \pi(\alpha)\, f_\alpha(b)/f(b) \\ &= \frac{\pi(\alpha)\, f_\alpha(\hat\beta)}{f(\hat\beta)} \frac{f_\alpha(b)}{f_\alpha(\hat\beta)} \frac{f(\hat\beta)}{f(b)}. \end{aligned} \tag{4.5}$$

But

$$\frac{f_\alpha(b)}{f_\alpha(\hat\beta)} = \exp\{(b-\hat\beta)^{\mathrm{T}}\alpha\} \frac{f_0(b)}{f_0(\hat\beta)}, \tag{4.6}$$

yielding

$$g(\alpha|b) = g(\alpha|\hat\beta) \exp\{(b-\hat\beta)^{\mathrm{T}}\alpha\} \frac{f_0(b)\, f(\hat\beta)}{f_0(\hat\beta)\, f(b)}. \tag{4.7}$$

The final factor does not depend on α and so must equal $\exp\{-\phi(b)\}$ in expression (4.3)–(4.4) for equation (4.7) to integrate to 1. □

In Sections 2 and 3, $g(\alpha|\hat\beta)$ was approximated by a discrete distribution putting weight p_i on α_i, say

$$\hat g(\alpha_i|\hat\beta) = p_i \qquad \text{for } i = 1, 2, \ldots, B, \tag{4.8}$$

$p_i = \pi_i R_i / \Sigma_1^B \pi_j R_j$ in bootstrap implementation (3.5)–(3.9), and $p_i = 1/B$ in the MCMC implementation (2.11) where the μ_i play the role of the α_i.

Substituting $\hat g(\alpha|\hat\beta)$ for $g(\alpha|\hat\beta)$ in expression (4.3) produces the *empirical posterior family* $\hat{\mathcal{G}}$. Define

$$W_i(b) = \exp\{(b-\hat\beta)^{\mathrm{T}}\alpha_i\}. \tag{4.9}$$

Then $\hat{\mathcal{G}}$ can be expressed as

$$\hat{\mathcal{G}} : \left\{ \hat g(\alpha_i|b) = W_i(b)\, p_i \Big/ \sum_{j=1}^{B} W_j(b)\, p_j \text{ for } i = 1, 2, \ldots, B \right\}, \tag{4.10}$$

$b \in \mathcal{R}^p$, i.e. the discrete distribution putting weight proportional to $W_i(b)\, p_i$ on α_i. (Note that $\hat{\mathcal{G}}$ differs from the empirical exponential family in section 6 of Efron (2012).)

We can now execute the 'straightforward bootstrap approach' (4.1) without much additional computation. The kth bootstrap replication $\hat\theta_k^* = \hat E\{t(\alpha)|\hat b_k^*\}$ is estimated from $\hat g(\alpha_i|\hat b_k^*)$, using the importance sampling formula, as

$$\hat\theta_k^* = \sum_{i=1}^{B} t_i\, W_i(\hat b_k^*)\, p_i \Big/ \sum_{i=1}^{B} W_i(\hat b_k^*)\, p_i \qquad t_i = t(\alpha_i). \tag{4.11}$$

Aside from step (4.1), usually comparatively inexpensive to carry out, we can obtain $\hat\theta_1^*, \hat\theta_2^*, \ldots, \hat\theta_K^*$ from just the original calculations for $\hat\theta = \Sigma t_i\, p_i$ and use the $\hat\theta_k^*$-values to construct a bootstrap confidence interval. (In particular, there is no need for new MCMC simulations for each new $\hat b_k^*$.)

Section 6 of Efron (2012) carries out this program under the rubric 'bootstrap after bootstrap'. It involves, however, some numerical peril: the weighting factors $W_i(\hat b_k^*)$ can easily blow up, destabilizing the estimates $\hat\theta_k^*$. The peril can be avoided by *local resampling*, i.e. by considering alternate data values b very close to the actual $\hat\beta$, rather than full bootstrap resamples as in expression (4.1).

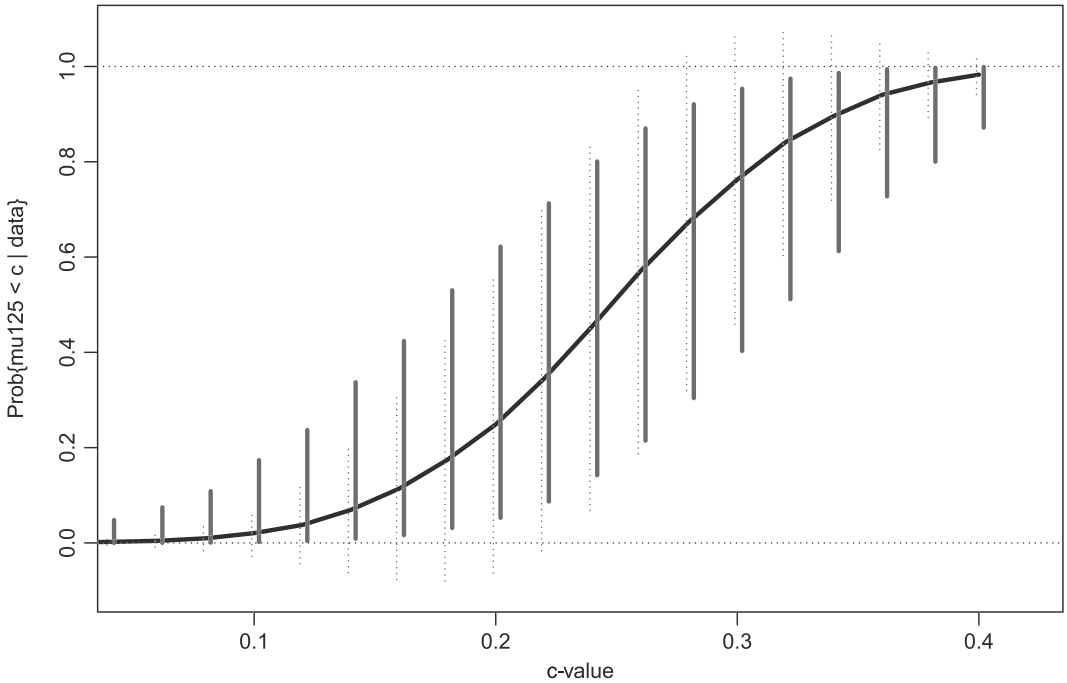

Fig. 3. Vertical bars are 68% central ABC limits for patient 125's posterior CDF in Fig. 1: they remain within the feasible interval [0, 1], unlike Fig. 1's standard deviation bars (⋮)

This suggests the 'approximate bootstrap confidence (ABC)' (DiCiccio and Efron, 1992) system of confidence intervals (not to be confused with 'approximate Bayesian computation', as in Fearnhead and Prangle (2012)). The ABC algorithm approximates full bootstrap confidence intervals by using only a small number of resamples b in the immediate neighbourhood of the observed sufficient statistic $\hat{\beta}$.

Fig. 3 shows again the posterior CDF from Fig. 1 for γ_{125}, the progression parameter for patient 125 in the diabetes study. The bold vertical bars indicate ABC 68% central frequentist confidence limits for the Bayes posterior CDF values. Now the confidence limits stay within [0, 1]. (95% limits are much wider, nearly filling the interval [0, 1] for some values of c, indicating that perhaps we are asking too much of the diabetes data set.) Remark 7 of Section 6 discusses the details of the ABC calculations.

Standard confidence intervals, say $\hat{\theta} \pm \widehat{\mathrm{sd}}$ for approximate 68% coverage, require only the original point estimate $\hat{\theta}$ and its accuracy estimate $\widehat{\mathrm{sd}}$, which in our case is what the general accuracy formula efficiently provides. The standard intervals are 'first order accurate', with their actual coverage probabilities converging to the nominal value at rate $n^{-1/2}$ as the sample size n grows large.

The ABC algorithm provides *second-order accuracy*, i.e. coverage errors approaching 0 at rate n^{-1}. This is more than a theoretical nicety. As the examples in DiCiccio and Efron (1992) showed, the ABC intervals often come close to exact small sample intervals when the latter exist. Three corrections are made to the standard intervals: for bias, for acceleration (i.e. changes in standard deviation between the interval end points) and for non-normality. The algorithm depends on exponential family structure, provided by $\hat{\mathcal{G}}$ the empirical posterior family (4.10), and a smoothly varying point estimate.

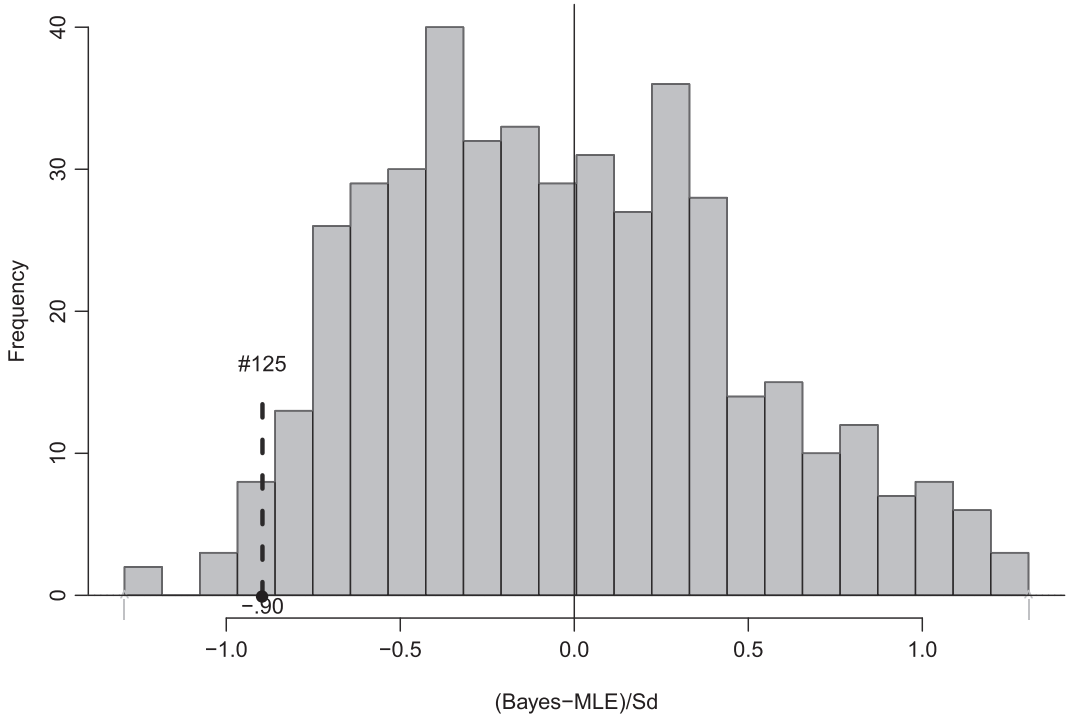

Fig. 4. Relative differences (4.14) for the 442 diabetes patients, Park and Casella prior (2.23): Bayes estimate minus the MLE, divided by the MLE standard deviation

In our situation the point estimate is the empirical posterior expectation (4.11) of $t(\alpha)$ given sufficient statistic b, say $\hat{\theta} = s(b)$,

$$\hat{\theta} = s(b) = \sum_{i=1}^{B} t_i \, W_i(b) \, p_i \Big/ \sum_{i=1}^{B} W_i(b) \, p_i. \tag{4.12}$$

For b near $\hat{\beta}$, the values that are explored in the ABC algorithm, the smoothness of the kernel $W_i(b)$ (4.9) makes $s(b)$ smoothly differentiable.

What parameter is the intended target of the ABC intervals? The answer, from DiCiccio and Efron (1992), is $\theta = s(\beta)$, the value of $s(b)$ if sufficient statistic b equals its expectation β. It is *not* $\gamma = t(\alpha)$, the true value of the parameter of special interest.

ABC's output includes *bias*, an assessment of the bias of $\hat{\theta} = s(\hat{\beta})$ as an estimator of θ, not as an estimate of γ. The more interesting quantity *definitional bias*,

$$\theta - \gamma = E\{t(\hat{\alpha}) \,|\, \hat{\beta} = \beta\} - t(\alpha), \tag{4.13}$$

depends on the prior $\pi(\alpha)$. It seems reasonable to ask that an uninformative prior should not produce large definitional biases. The parameter γ_{125} (2.29) has MLE 0.316 ± 0.076, compared with Bayes estimate and frequentist standard deviation 0.248 ± 0.071, giving a relative difference of

$$\hat{\delta} = \frac{\hat{\theta} - \hat{\gamma}}{\mathrm{sd}(\hat{\gamma})} = \frac{0.248 - 0.316}{0.076} = -0.90. \tag{4.14}$$

In other words, the Park and Casella (2008) prior (2.23) shifts the estimate for patient 125 about 0.9 standard deviations downwards, quite a substantial effect.

Fig. 4 shows the relative difference estimates for all 442 diabetes patients. Most of the $\hat{\delta}$s are less extreme than that for patient 125. Even though prior (2.23) looks like a strong shrinker, and not at all uninformative, its effects on the patient estimates are mostly moderate.

5. Hierarchical and empirical Bayes accuracy

Modern scientific technology excels at the simultaneous execution of thousands, and more, parallel investigations, the iconic example being microarray studies of genetic activity. Both hierarchical and empirical Bayes methods provide natural statistical tools for analysing large parallel data sets. This section compares the accuracy of the two methods, providing some intuition about why, often, there is not much difference.

A typical hierarchical model begins with a *hyperprior* $\pi(\alpha)$ providing a *hyperparameter* α, which determines a prior density $g_\alpha(\delta)$; N realizations are generated from $g_\alpha(\cdot)$, say

$$\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_k, \ldots, \delta_N); \tag{5.1}$$

finally, each parameter δ_k provides an observation z_k according to density $h_{\delta_k}(z_k)$, yielding a vector $\mathbf{z}$ of N observations,

$$\mathbf{z} = (z_1, z_2, \ldots, z_k, \ldots, z_N). \tag{5.2}$$

The functional forms $\pi(\cdot)$, $g_\alpha(\cdot)$ and $h_\delta(\cdot)$ are known, but not the values of α and $\boldsymbol{\delta}$. Here we shall assume that the pairs (δ_k, z_k) are generated independently for $k = 1, 2, \ldots, N$. We wish to estimate the parameter $\boldsymbol{\delta}$ from the observations $\mathbf{z}$.

If α were known then Bayes theorem would directly provide the conditional distribution of δ_k given z_k:

$$g_\alpha(\delta_k|z_k) = g_\alpha(\delta_k)\, h_{\delta_k}(z_k)/f_\alpha(z_k), \tag{5.3}$$

where $f_\alpha(z_k)$ is the marginal density of z_k given α,

$$f_\alpha(z_k) = \int g_\alpha(\delta)\, h_\delta(z_k)\, \mathrm{d}\delta. \tag{5.4}$$

The empirical Bayes approach estimates the unknown value of α from the observed vector $\mathbf{z}$, often by marginal maximum likelihood:

$$\hat{\alpha} = \arg\max_\alpha \left\{ \prod_{i=1}^{N} f_\alpha(z_k) \right\}, \tag{5.5}$$

and then infers the individual δ_ks according to $g_{\hat{\alpha}}(\delta_k|z_k)$. Hierarchical Bayes inference aims instead for the full posterior distribution of δ_k given all the observations $\mathbf{z}$:

$$g(\delta_k|\mathbf{z}) = \int g_\alpha(\delta_k|z_k)\, \pi(\alpha|\mathbf{z})\, \mathrm{d}\alpha. \tag{5.6}$$

We shall employ the general accuracy formula to compare the frequentist variability of the two approaches. In the example that follows, all of the calculations can be carried out in terms of the marginal densities $f_\alpha(\cdot)$, rendering it unnecessary to specify the prior densities $g_\alpha(\cdot)$.

As a working example we consider the prostate cancer microarray data (Singh *et al.*, 2002). Each of 102 men, 52 prostate cancer patients and 50 controls, has had the activity of $N = 6033$ genes measured, as discussed in Section 5 of Efron (2012). A test statistic z_k comparing cancer

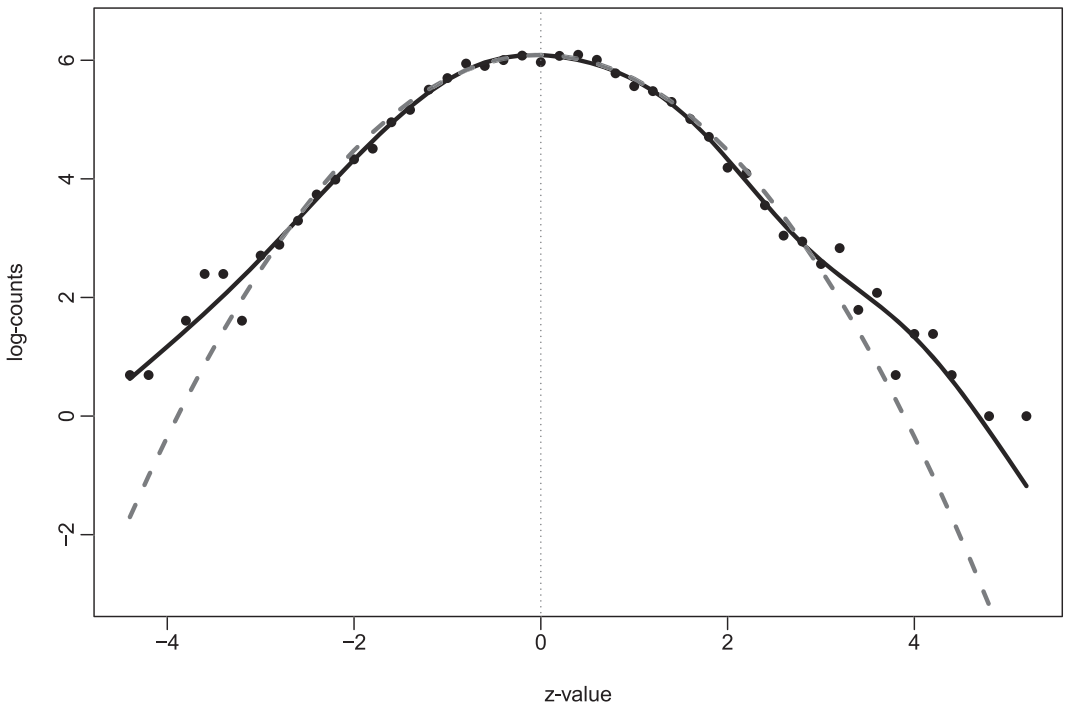

Fig. 5. Prostate cancer data (•, log-counts for 49 bins (5.8)–(5.9)): the quadratic curve (– – –) would fit the log-counts if all the genes were null, $\delta_k = 0$ in distribution (5.7); the eighth-degree polynomial (———) gives a much better fit, indicating that some genes have large effect sizes

patients with controls has been calculated for each gene, which we shall assume here follows a normal translation model

$$z_k \sim \mathcal{N}(\delta_k, 1), \tag{5.7}$$

where δ_k is gene k's *effect size* (so $h_\delta(z)$ in equations (5.3) and (5.4) is the normal kernel $\phi(z-\delta) = \exp\{-(z-\delta)^2/2\}/\sqrt{(2\pi)}$). 'Null' genes have $\delta_k = 0$ and $z_k \sim \mathcal{N}(0,1)$, but of course the investigators were looking for non-null genes: those having large δ_k-values, either positive or negative.

Binning the data simplifies the Bayes and empirical Bayes analyses. For Fig. 5 the data have been put into $J = 49$ bins $\mathcal{Z}_j$, each of width 0.2, with centres c_j,

$$c_j = -4.4, -4.2, \ldots, 5.2. \tag{5.8}$$

Let y_j be the count in bin $\mathcal{Z}_j$:

$$y_j = \#\{z_k \in \mathcal{Z}_j\}. \tag{5.9}$$

The dots in Fig. 5 are the log-counts $\log(y_j)$. The broken quadratic curve would give a good fit to the dots if all the genes were null, but it is obviously deficient in the tails, suggesting some large effect sizes.

An eighth-degree polynomial (the full curve) provided a good fit to the data. It was obtained from a Poisson regression generalized linear model. The counts y_j (5.9) were assumed to be independent Poisson variates,

$$y_j \overset{\text{ind}}{\sim} \text{Poi}(\mu_j), \qquad j = 1, 2, \ldots, J = 49, \tag{5.10}$$

with

$$\mu_j = E_\alpha(y_j) = \exp\{\mathbf{x}(c_j)\alpha\}. \tag{5.11}$$

Here $\mathbf{x}(c_j)$ is the nine-dimensional row vector

$$\mathbf{x}(c_j) = (1, c_j, c_j^2, \ldots, c_j^8), \tag{5.12}$$

the c_j being the bin centres (5.8), and α is an unknown parameter vector, $\alpha \in \mathcal{R}^9$. There is a small loss of information in going from the full data vector $\mathbf{z}$ to the binned counts that we shall ignore here.

Model (5.10)–(5.12) is a nine-parameter exponential family $f_\alpha(\hat{\beta})$ (3.1) with α the natural parameter vector. Its sufficient statistic is

$$\hat{\beta} = \mathbf{X}^{\mathrm{T}}\mathbf{y}, \tag{5.13}$$

where $\mathbf{X}$ is the 49×9 matrix having jth row $\mathbf{x}(c_j)$, and $\mathbf{y}$ is the 49-vector of counts; $\hat{\beta}$ has covariance matrix

$$V_\alpha = \mathbf{X}^{\mathrm{T}} \operatorname{diag}(\boldsymbol{\mu}_\alpha)\mathbf{X}, \tag{5.14}$$

$\operatorname{diag}(\boldsymbol{\mu}_\alpha)$ the diagonal matrix with diagonal elements (5.11).

We are now ready to apply the accuracy formula in the exponential family form of theorem 2, expression (3.3). A notable feature of this example is that the parameter of interest $t(\alpha)$ is itself a posterior expectation: let $\tau(\delta)$ be an 'interesting function' of an individual parameter δ in equation (5.1), for instance the indicator of whether or not $\delta = 0$:

$$\tau(\delta) = I_0(\delta). \tag{5.15}$$

Letting (δ_0, z_0) represent a hypothetical (parameter, observation) pair, we define $t(\alpha)$ to be the conditional expectation of $\tau(\delta_0)$ given z_0, α and the sufficient statistic $\hat{\beta}$,

$$t(\alpha) = E\{\tau(\delta_0)|z_0, \alpha, \hat{\beta}\}. \tag{5.16}$$

In the prostate cancer study, for example, with $\tau(\delta) = I_0(\delta)$ and $z_0 = 3$, $t(\alpha)$ is the conditional probability of a gene being null given a z-value of 3. However, α is unobserved and $t(\alpha)$ must be inferred. The hierarchical Bayes estimate is

$$\hat{\theta} = E\{t(\alpha)|\hat{\beta}\} = E\{\tau(\delta_0)|z_0, \hat{\beta}\}, \tag{5.17}$$

compared with the empirical Bayes MLE estimate $t(\hat{\alpha})$. (Note that we now require three levels of parameter definition: in addition to $\hat{\theta}$ being the posterior expectation of $t(\alpha)$ (5.17), $t(\alpha)$ itself is the posterior expectation of $\tau(\delta_0)$ (5.16).)

The hyperprior $\pi(\alpha)$ is usually taken to be uninformative in hierarchical Bayes applications, making them good candidates for the bootstrap implementation of Section 3. Let $\hat{\alpha}$ be the MLE of hyperparameter α, which is obtained in the prostate cancer study by Poisson regression from model (5.10)–(5.12), `glm(y~X, poisson)$coef` in language R. From $\hat{\alpha}$ we obtain parametric bootstrap samples $\mathbf{y}_i^*$, $i = 1, 2, \ldots, B$:

$$y_{ij}^* \overset{\text{ind}}{\sim} \operatorname{Poi}(\hat{\mu}_j), \qquad j = 1, 2, \ldots, J, \tag{5.18}$$

where $\hat{\mu}_j = \exp\{\mathbf{x}(c_j)\hat{\alpha}\}$. The $\mathbf{y}_i^*$-vector yields β_i and α_i, expressions (3.5) and (3.6): $\beta_i = \mathbf{X}^{\mathrm{T}}\mathbf{y}^*$ and $\alpha_i = \texttt{glm}(\texttt{y}^* \sim \texttt{X, poisson})\$\texttt{coef}$.

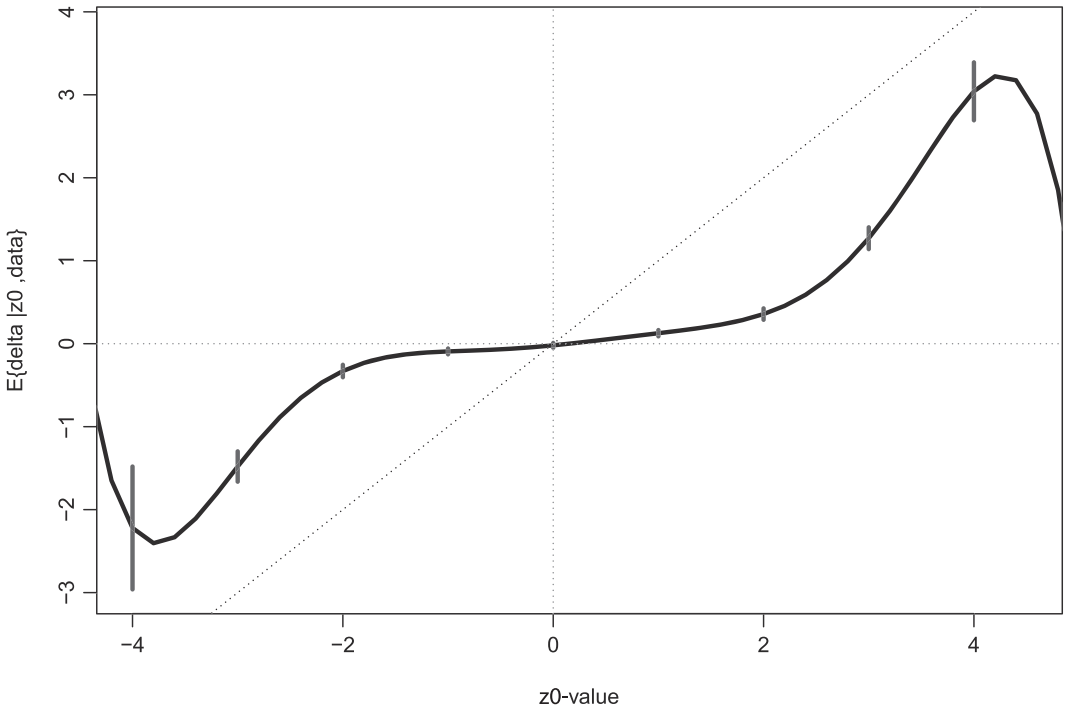

Fig. 6. Hierarchical Bayes estimate $\hat{\theta} = E(\delta_0|z_0, \hat{\beta})$ as a function of z_0 for the prostate cancer study data, calculated from $B = 4000$ parametric bootstrap samples (5.18) (|, ± 1 frequentist standard deviation $\widehat{\text{sd}}$ (3.11))

If for convenience we take $\pi(\alpha)$ to be Jeffreys prior, then the weights $\pi_i R_i$ in expression (3.7) become

$$\pi_i R_i = \exp\{\Delta(\alpha_i)\} \tag{5.19}$$

where, for Poisson regression, the half-deviance difference $\Delta(\alpha_i)$ is

$$\Delta_i = (\alpha_i - \hat{\alpha})^{\mathrm{T}}(\beta_i + \hat{\beta}) - 2\sum_{j=1}^{J}(\mu_{ij} - \hat{\mu}_j), \tag{5.20}$$

$\hat{\mu}_j = \exp\{\mathbf{x}(c_j)\hat{\alpha}\}$ and $\mu_{ij} = \exp\{\mathbf{x}(c_j)\alpha_i\}$ (Efron (2012), section 5). Letting t_i be the conditional expectation (5.16),

$$t(\alpha_i) = E\{\tau(\delta_0)|z_0, \alpha_i, \hat{\beta}\}, \tag{5.21}$$

the hierarchical Bayes estimate $\hat{\theta}$ (5.17) is

$$\hat{\theta} = \sum_{i=1}^{B} p_i t_i \qquad p_i = \exp(\Delta_i) \Big/ \sum_{k=1}^{B} c^{\Delta_k} \tag{5.22}$$

and has frequentist standard deviation $\widehat{\text{sd}}$ (3.11) from theorem 2.

Fig. 6 applies to the prostate cancer data, taking $\tau(\delta)$, the function of interest, to be δ itself, i.e. the hierarchical Bayes estimate (5.17) is

$$\hat{\theta} = E(\delta_0|z_0, \hat{\beta}), \tag{5.23}$$

Table 3. Comparison of hierarchical and empirical Bayes estimates for expected effect sizes in the prostate cancer study†

	Results for the following values of z_0*:*								
	−4	*−3*	*−2*	*−1*	*0*	*1*	*2*	*3*	*4*
1 Bayes estimate	−2.221	−1.480	−0.329	−0.093	−0.020	0.127	0.357	1.271	3.042
2 Empirical Bayes estimate	−2.217	−1.478	−0.331	−0.092	−0.020	0.126	0.360	1.266	3.021
3 Bayes standard deviation	0.756	0.183	0.074	0.036	0.030	0.039	0.071	0.131	0.336
4 Bayes frequentist standard deviation	0.740	0.183	0.075	0.035	0.029	0.038	0.068	0.131	0.349
5 Empirical Bayes standard deviation	0.878	0.187	0.074	0.037	0.030	0.039	0.072	0.139	0.386

†Row 1, Bayes estimate $\hat{\theta}$ (5.23); row 2, empirical Bayes estimate $E(\delta_0|z_0, \alpha = \hat{\alpha}, \hat{\beta})$; row 3, Bayes posterior standard deviation $\{\Sigma p_i(t_i - \hat{\theta})^2\}^{1/2}$; row 4, frequentist standard deviation of $\hat{\theta}$ (3.11); row 5, bootstrap standard deviation (5.25).

the posterior expected effect size for a gene having $z = z_0$. The calculations assume Poisson regression model (5.10)–(5.12), beginning with Jeffreys prior $\pi(\alpha)$. $B = 4000$ bootstrap samples (5.18) provided the Bayesian estimates, as in equations (5.19)–(5.22). (Tweedie's formula (Efron, 2011) says that $\hat{\theta}$ in equation (5.23) equals $z_0 + \mathrm{d}/\mathrm{d}z \log\{f_{\hat{\alpha}}(z)\}|_{z_0}$, again allowing us to avoid explicit characterization of the prior distributions $g_\alpha(\cdot)$.)

The bold curve in Fig. 6 shows $\hat{\theta}$ as a function of z_0. It stays near zero for z_0 in $[-2, 2)$, suggesting nullity for genes having small z-values, and then swings away from the horizontal axis, indicating non-null effect sizes for large $|z_0|$, but always with strong 'regression to the mean' behaviour: $|\hat{\theta}| < |z_0|$. The vertical bars span ± 1 frequentist standard deviation $\widehat{\mathrm{sd}}$ (3.11).

There was very little difference between the hierarchical and empirical Bayes results. The graph of the empirical Bayes estimates

$$t(\hat{\alpha}) = E(\delta_0|z_0, \alpha = \hat{\alpha}, \hat{\beta}) \tag{5.24}$$

follows the curve in Fig. 6 to within the line width. Table 3 gives numerical comparisons for $z_0 = -4, -3, \ldots, 4$. The estimated standard deviations for the empirical Bayes estimates (row 5) are a little bigger than those in row 3 for hierarchical Bayes estimates, but that may just reflect the fact that the former are full bootstrap estimates whereas the latter are delta method standard deviations.

Particularly striking is the agreement between the frequentist standard deviation estimates for $\hat{\theta}$ (3.11) (row 4) and the posterior Bayes standard deviation estimates (row 3). This is the predicted *asymptotic* behaviour (Berger (1985), section 4.7.8) if the effect of the prior distribution has indeed been swamped by the data. It cannot be assumed, though, that agreement would hold for estimates other than equation (5.23).

The empirical Bayes estimate $t(\hat{\alpha}) = E(\delta_0|z_0, \alpha = \hat{\alpha}, \hat{\beta})$ had its standard deviation $\overline{\mathrm{sd}}$ (row 5 of Table 3) calculated directly from its bootstrap replications,

$$\overline{\mathrm{sd}} = \left\{ \sum_1^B (t_i - \bar{t})^2 / B \right\}^{1/2} \qquad \bar{t} = \sum_1^B t_i / B, \tag{5.25}$$

compared with the Bayes posterior standard deviation (row 3)

$$\widehat{\mathrm{sd}} = \left\{ \sum_1^B p_i (t_i - \hat{\theta})^2 \right\}^{1/2} \qquad \hat{\theta} = \sum_1^B p_i t_i. \tag{5.26}$$

Table 4. Polynomial model selection for the prostate cancer study data†

		Results (%) for the following values of m:				
		4	*5*	*6*	*7*	*8*
1	Bootstrap %	32	10	5	1	51
2	Bayes expected	36	12	5	2	45
3	Frequentist standard deviation	±32	±16	±8	±3	±40

†Row 1, raw bootstrap proportions for best polynomial fit, Akaike information criterion; row 2, corresponding Bayes posterior probabilities, Jeffreys prior; row 3, frequentist standard deviations for the Bayes estimates.

(See remark 8 of Section 6 concerning the calculation of t_i.) The difference comes from weighting the B bootstrap replications t_i according to p_i (3.7), rather than equally. Lemma 3 of Efron (2012) shows that the discrepancy, which is small in Table 3, depends on the empirical correlation between p_i and t_i.

There is a similar relationship between rows 4 and 5 of Table 3. Remark 9 shows that $\overline{\text{sd}}$ (row 5) is approximated by

$$\overline{\text{sd}} \doteq (\overline{\text{cov}}^{\text{T}} V_{\hat{\alpha}} \overline{\text{cov}})^{1/2}, \tag{5.27}$$

where $\overline{\text{cov}}$ is the unweighted bootstrap covariance between α_i and t_i:

$$\overline{\text{cov}} = \sum_1^B (\alpha_i - \bar{\alpha})(t_i - \bar{t})/B \qquad \bar{\alpha} = \sum_1^B \alpha_i/B. \tag{5.28}$$

This compares with the weighted version (3.10)–(3.11) of row 4. Weighting did not matter much in Table 3, leaving the three standard deviations more alike than different.

The eighth-degree polynomial fit that was used in Fig. 5 might be excessive. For each of the $B = 4000$ bootstrap samples $\mathbf{y}_i^*$, the 'best' polynomial degree m_i^* was selected according to the Akaike information criterion, as detailed in section 5 of Efron (2012). Only degrees $m = 0$–8 were considered. The top row of Table 4 shows that 32% of the 4000 bootstrap samples gave $m_i^* = 4$, compared with 51% for $m_i^* = 8$. (None of the samples had m_i^* less than 4.)

Let $t_i^{(m)}$ be the indicator for model m selection:

$$t_i^{(m)} = \begin{cases} 1 & \text{if } m_i^* = m, \\ 0 & \text{if } m_i^* \neq m. \end{cases} \tag{5.29}$$

Then

$$\hat{\theta}^{(m)} = \sum_{i=1}^B p_i t_i^{(m)} \tag{5.30}$$

is the posterior probability of the region $\mathcal{R}^{(m)}$ in the space of possible α-vectors where degree m is best; for instance, $\hat{\theta}^{(4)}$ equals 36% in row 2.

We can apply theorem 2, expression (3.11), to obtain frequentist standard deviations for the $\hat{\theta}^{(m)}$. These are shown in row 3 of Table 4. The results are discouraging, with $\hat{\theta}^{(4)} = 36\%$ having $\widehat{\text{sd}} = 32\%$ and so on. (These numbers differ from those in Table 2 of Efron (2012), where the standard deviations were assessed by the potentially perilous 'bootstrap-after-bootstrap' method.) There was a strong negative frequentist correlation of -0.84 between $\hat{\theta}^{(4)}$ and $\hat{\theta}^{(8)}$ (using expression (2.20)). All of this suggests that the MLE $\hat{\alpha}$ lies near the boundary between

$\mathcal{R}^{(4)}$ and $\mathcal{R}^{(8)}$, but not near the other regions. Bayesian model selection, of the limited type that was considered above, is frequentistically unstable for the prostate cancer data.

6. Remarks

This section presents remarks, details and extensions of the previous material.

6.1. Remark 1: relationship of Bayes and frequentist standard deviations

In several of our examples the posterior Bayes estimate $\hat{\theta}$ had its posterior standard deviation $\widehat{\mathrm{sd}}_{\mathrm{Bayes}}$ quite close to $\widehat{\mathrm{sd}}_{\mathrm{freq}}$, the frequentist standard deviation. Why this might happen, or might not, is easy to understand in the diabetes data example (2.29)–(2.30).

Let $\tilde{\boldsymbol{\alpha}}$ be the $10\,000 \times 10$ matrix with ith row $\alpha_i - \bar{\alpha}$, so

$$\Sigma_\alpha = \tilde{\boldsymbol{\alpha}}^{\mathrm{T}} \tilde{\boldsymbol{\alpha}} / B \qquad B = 10\,000 \tag{6.1}$$

is the empirical covariance matrix of the α_i-vectors. For any fixed row vector $\mathbf{x}$ we define as our parameter of special interest $\gamma_{\mathbf{x}} = \mathbf{x}\alpha$ ($\mathbf{x} = \mathbf{x}_{125}$ in expression (2.29)). Each α_i gives $t_i = \mathbf{x}\alpha_i$, with average $\bar{t} = \mathbf{x}\bar{\alpha}$. The vector $\tilde{\mathbf{t}}$ of centred values $\tilde{t}_i = t_i - \bar{t}$ is given by

$$\tilde{\mathbf{t}} = \tilde{\boldsymbol{\alpha}} \mathbf{x}^{\mathrm{T}}. \tag{6.2}$$

Then

$$\widehat{\mathrm{sd}}^2_{\mathrm{Bayes}} = \sum_1^B \tilde{\mathbf{t}}^2 / B = \mathbf{x} \Sigma_\alpha \mathbf{x}^{\mathrm{T}}. \tag{6.3}$$

Also, from equation (2.13),

$$\widehat{\mathrm{cov}}^{\mathrm{T}} = \tilde{\mathbf{t}} \tilde{\boldsymbol{\alpha}} / B = \mathbf{x} \Sigma_\alpha, \tag{6.4}$$

yielding

$$\widehat{\mathrm{sd}}^2_{\mathrm{freq}} = \mathbf{x} \Sigma_\alpha G \Sigma_\alpha \mathbf{x}^{\mathrm{T}} \tag{6.5}$$

from expression (2.28).

The variance ratio $\mathrm{rat}(\mathbf{x})$ is

$$\mathrm{rat}(\mathbf{x}) = \left(\frac{\widehat{\mathrm{sd}}_{\mathrm{freq}}}{\widehat{\mathrm{sd}}_{\mathrm{Bayes}}} \right)^2 = \frac{\mathbf{x} \Sigma_\alpha G \Sigma_\alpha \mathbf{x}^{\mathrm{T}}}{\mathbf{x} \Sigma_\alpha \mathbf{x}^{\mathrm{T}}}. \tag{6.6}$$

Suppose that $H = \Sigma_\alpha^{1/2} G \Sigma_\alpha^{1/2}$ has spectral decomposition $H = \Gamma \mathbf{d} \Gamma^{\mathrm{T}}$, with $\mathbf{d}$ the diagonal matrix of eigenvalues. Then equation (6.6) reduces to

$$\mathrm{rat}(\mathbf{x}) = \sum_1^p d_i v_i^2 \Big/ \sum v_i^2 \qquad (\mathbf{v} = \mathbf{x} \Sigma_\alpha^{1/2} \Gamma) \tag{6.7}$$

Table 5 shows the eigenvalues d_i. We see that $\mathrm{rat}(\mathbf{x})$ could vary from 1.014 down to 0.098. For the 442 diabetes patients, $\mathrm{rat}(\mathbf{x}_i)$ ranged from 0.991 to 0.670, averaging 0.903; $\mathrm{rat}(\mathbf{x}_{125}) = 0.962$ was near the high end. A spherically uniform choice of $\mathbf{v}$ in expression (6.7) would yield an average $\mathrm{rat}(\mathbf{x})$ of 0.800.

The fact that the eigenvalues in Table 5 are mostly less than 1 relates to the Park and Casella (2008) prior (2.23). A flat prior for model (2.22) has $\mathrm{cov}(\boldsymbol{\alpha}) = G^{-1}$, giving $H = I$ and eigenvalues $d_i = 1$ in equation (6.7). The Park and Casella prior (2.23) is a 'shrinker', making Σ_α and H less than I.

Table 5. Eigenvalue d_i for the variance ratio rat(**x**) (6.7)

d_i	1.014	1.009	0.986	0.976	0.961	0.944	0.822	0.710	0.482	0.098

A more general but less transparent formula for $(\widehat{\mathrm{sd}}_{\mathrm{freq}}/\widehat{\mathrm{sd}}_{\mathrm{Bayes}})^2$ is available for possibly nonlinear parameters $t(\alpha)$. As before, let p_i be the weight on α_i, with p_i equalling $1/B$ or expression (3.7) in Sections 2 and 3 respectively, giving $\bar{t}=\Sigma\, p_i t_i$ and $\bar{\alpha}=\Sigma\, p_i\alpha_i$. Define $s_i=\sqrt{p_i}(t_i-\bar{t})$ and matrix M,

$$M=\mathrm{diag}(p_i^{1/2})\tilde{\boldsymbol{\alpha}}V_{\hat{\alpha}}\tilde{\boldsymbol{\alpha}}^{\mathrm{T}}\,\mathrm{diag}(p_i^{1/2}), \tag{6.8}$$

where $\tilde{\boldsymbol{\alpha}}$ has rows $\alpha_i-\bar{\alpha}$ and $V_{\hat{\alpha}}$ is as in expression (3.11). The spectral decomposition $M=\Gamma\mathbf{d}\Gamma'$ has $p=\mathrm{rank}(\tilde{\boldsymbol{\alpha}})$ non-zero eigenvalues d_i, with corresponding eigenvectors Γ_i, giving, after straightforward calculations,

$$\left(\frac{\widehat{\mathrm{sd}}_{\mathrm{freq}}}{\widehat{\mathrm{sd}}_{\mathrm{Bayes}}}\right)^2=\frac{\sum_1^p d_i s_i^2}{\sum_1^p s_i^2}\qquad s_i=\mathbf{s}^{\mathrm{T}}\Gamma_i \tag{6.9}$$

for $\hat{\theta}=\Sigma\, p_i t_i$; the ratio can range from a high of d_1 to a low of d_p, depending on how $t(\alpha)$ aligns with the eigenvectors of M.

6.2. Remark 2: a computational verification of lemma 1

Working directly with the implementation values μ_i, α_i and t_i (2.11)–(2.12), we can verify lemma 1 in the form in which it is actually used computationally. For $\tilde{x}$ any point in the sample space of the sufficient statistic, define

$$W_\mu(\tilde{x})=f_\mu(\tilde{x})/f_\mu(x), \tag{6.10}$$

with x the observed statistic. Letting $\tilde{x}=x+\mathrm{d}x$ with $\mathrm{d}x\to 0$,

$$W_\mu(\tilde{x})=\frac{f_\mu(x)+f'_\mu(x)\,\mathrm{d}x+o(\mathrm{d}x)}{f_\mu(x)}=1+\alpha_x(\mu)\,\mathrm{d}x+r(x) \tag{6.11}$$

where the remainder $r(x)=o(\mathrm{d}x)/f_\mu(x)$. Here we are assuming that $f_\mu(x)$ has continuous gradient $f'_\mu(\tilde{x})$ in a neighbourhood of x, and that $f_\mu(x)>0$.

The importance sampling estimate of $E\{t(\mu)|\tilde{x}\}$ is

$$\begin{aligned}\hat{\theta}(\tilde{x})&=\sum_{i=1}^B t_i\,W_i(\tilde{x})\Big/\sum_{i=1}^B W_i(\tilde{x})\\&=\sum t_i(1+\alpha_i\,\mathrm{d}x+r_i)\Big/\sum(1+\alpha_i\,\mathrm{d}x+r_i),\end{aligned} \tag{6.12}$$

with $W_i=W_{\mu_i}(\tilde{x})$, $\alpha_i=\alpha_x(\mu_i)$ and $r_i=o_i(\mathrm{d}x)/f_{\mu_i}(x)$. Denoting $\bar{t}=\Sigma t_i/B$, $\overline{t\alpha}=\Sigma t_i\alpha_i/B$, etc., equation (6.12) gives

$$\hat{\theta}(x+\mathrm{d}x)=\frac{\bar{t}\{1+(\overline{t\alpha}/\bar{t})\,\mathrm{d}x+\overline{tr}/\bar{t}\}}{1+\bar{\alpha}\,\mathrm{d}x+\bar{r}}. \tag{6.13}$$

Since $\bar{t} = \hat{\theta}(x)$ and $\overline{tr}$ and $\bar{r}$ are $o(\mathrm{d}x)$, letting $\mathrm{d}x \to 0$ yields

$$\begin{aligned}\hat{\theta}(x + \mathrm{d}x) &= \hat{\theta} + (\overline{t\alpha} - \bar{t}\bar{\alpha})\,\mathrm{d}x + o(\mathrm{d}x) \\ &= \hat{\theta} + \widehat{\mathrm{cov}}\,\mathrm{d}x + o(\mathrm{d}x),\end{aligned} \tag{6.14}$$

with $\widehat{\mathrm{cov}}$ as in equation (2.13). This verifies lemma 1 as employed in the computational form of theorem 1: $\widehat{\mathrm{sd}} = (\widehat{\mathrm{cov}}^{\mathrm{T}} V_{\hat{\mu}} \widehat{\mathrm{cov}})^{1/2}$ (3.11).

6.3. Remark 3: an alternative form of lemma 1

Lemma 1 assumes the computational form $\nabla_{\hat{\beta}}\hat{\theta} = \widehat{\mathrm{cov}}(t, \alpha)$ (3.10) in an exponential family (3.1). Defining

$$O_i = Q_i/\bar{Q} - P_i/\bar{P} \tag{6.15}$$

as in equation (3.12), an equivalent expression for $\widehat{\mathrm{cov}}$ turns out to be

$$\widehat{\mathrm{cov}} = \hat{\theta}\,\mathrm{cov}_*(O, \alpha), \tag{6.16}$$

where cov_* is the usual *unweighted* bootstrap covariance

$$\mathrm{cov}_* = \sum_{i=1}^{B} (\alpha_i - \bar{\alpha}) O_i / B \qquad \bar{\alpha} = \sum_{1}^{B} \alpha_i / B. \tag{6.17}$$

(Note that $\bar{O} = 0$.) This leads to a convenient formula for the frequentist coefficient of variation $\widehat{\mathrm{sd}}/|\hat{\theta}|$ of $\hat{\theta}$,

$$\widehat{\mathrm{cv}} = (\mathrm{cov}_*^{\mathrm{T}} V_{\hat{\alpha}}\,\mathrm{cov}_*)^{1/2}, \tag{6.18}$$

compared with the *internal* coefficient of variation $\mathrm{sd}_*(O)/\sqrt{B}$ (3.12).

6.4. Remark 4: bias correction for $\widehat{\mathrm{sd}}$

Monte Carlo calculation of $\widehat{\mathrm{sd}}$, either by MCMC or bootstrap methods, can be improved by a downward internal bias correction. Define $\breve{O}_i = \hat{\theta} O_i$ (6.15), $\breve{\alpha}_i = V_{\hat{\alpha}}^{1/2} \alpha_i$, and vector

$$C_B = \mathrm{cov}_*(\breve{O}, \breve{\alpha}) = \sum_{i=1}^{B} \breve{\alpha}_i \breve{O}_i / B. \tag{6.19}$$

Then formula (6.18) can be re-expressed as

$$\widehat{\mathrm{sd}}^2 = \|C_B\|^2. \tag{6.20}$$

Let C_∞ denote the limit of C_B as the number of parametric bootstrap replications $B \to \infty$. The last expression in equation (6.19) suggests that C_B has approximate bootstrap expectation and covariance

$$C_B \sim (C_\infty, D_B), \tag{6.21}$$

with D_B the component of covariance from stopping at B replications rather than going on to ∞. Combining expressions (6.20) and (6.21) gives

$$\widehat{\mathrm{sd}}^2 = \widehat{\mathrm{sd}}_\infty^2 + \mathrm{tr}(D_B) \tag{6.22}$$

($\widehat{\mathrm{sd}}_\infty$ being the ideal standard deviation estimate when $B \to \infty$), indicating an upward bias in $\widehat{\mathrm{sd}}$.

The bias-corrected standard deviation estimate for $\hat{\theta}$ is given by

$$\breve{\mathrm{sd}}^2 = \widehat{\mathrm{sd}}^2 - \mathrm{tr}(D_B). \tag{6.23}$$

Jackknife calculations provide a convenient estimate of $\mathrm{tr}(D_B)$: the B bootstrap replications are divided into J groups of B/J each (e.g. $J=20$); C_{Bj} is computed as in equation (6.19) but with the jth group of replications removed, giving the $J \times p$ matrix $\mathbf{C}$ with rows C_{Bj}; finally the $p \times p$ sample covariance matrix of $\mathbf{C}$ gives the estimate

$$\mathrm{tr}(D_B) = \frac{(J-1)^2}{J}\mathrm{tr}\{\mathrm{cov}(\mathbf{C})\}. \tag{6.24}$$

D_B decreases at rate $1/B$, and the large choices of B in our examples made the bias correction (6.23) insignificant.

6.5. Remark 5: binomial deviance difference

The binomial generalized linear model for the cell infusion data analysis (3.17)–(3.18) has half-deviance difference

$$\Delta = \sum_{j,k=1}^{5} ((\eta_{jk} - \hat{\eta}_{jk})(\xi_{jk} + \hat{\xi}_{jk}) - 2\log[\{1+\exp(\eta_{jk})\}/\{1+\exp(\hat{\eta}_{jk})\}]), \tag{6.25}$$

where $\eta_{jk} = \log\{\xi_{jk}/(1-\xi_{jk})\}$. Here we have suppressed subscript i.

6.6. Remark 6: a vector parameter example

The joint frequentist behaviour of the 0.90 credible interval end points $[0.292, 0.380]$ in Fig. 2 involved the vector parameter form (2.20) of the general accuracy formula, carried out by the bootstrap sampling method of Section 3.

With $I_c(\gamma)$ the indicator function of $\gamma \leqslant c$, we define the bivariate parameter replication $t_i = (I_{2.92}(\gamma_i), I_{3.80}(\gamma_i))$ for $i = 1, 2, \ldots, B = 2000$. Then $\widehat{\mathrm{cov}}$ (3.10) is a 2×2 matrix, as is $\widehat{\mathrm{var}}$ (3.11). The weighted bootstrap density $\hat{f}(\gamma)$ had numerical derivatives $(d_{\mathrm{lo}}, d_{\mathrm{up}}) = (0.466, 0.330)$ at the interval end points;

$$\begin{pmatrix} d_{\mathrm{lo}} & 0 \\ 0 & d_{\mathrm{up}} \end{pmatrix}^{-1} \widehat{\mathrm{var}} \begin{pmatrix} d_{\mathrm{lo}} & 0 \\ 0 & d_{\mathrm{up}} \end{pmatrix} = \begin{pmatrix} 0.0476 & 0.0678 \\ 0.0678 & 0.0968 \end{pmatrix} \tag{6.26}$$

is the usual delta method covariance matrix estimate for the end points, giving them frequentist standard deviations 0.218 and 0.311, and correlation 0.999.

6.7. Remark 7: approximate bootstrap confidence calculations for the diabetes data

The `abc` algorithm (DiCiccio and Efron, 1992) provides second-order-accurate confidence intervals for scalar parameters $\theta = T(\beta)$ in p-parameter exponential families (3.1). It does this by recomputing the MLE $\hat{\theta} = T(\hat{\beta})$ for values of b near $\hat{\beta}$ (only $4p+4$ recomputations are needed), calculating $2p+2$ numerical second derivatives, and using these to make second-order adjustments to the standard intervals $\hat{\theta} \pm c\,\widehat{\mathrm{sd}}$. An R version of `abc` is available from the author.

The full bars in Fig. 3 are ABC intervals for the point estimates

$$\hat{\theta}_c = \widehat{\Pr}(\gamma_{125} \leqslant c | \hat{\beta}), \tag{6.27}$$

(2.29). Here $\hat{G}$ (4.10) was the p-parameter exponential family, $p=10$, with α_i (2.26) the $B=10\,000$ MCMC vectors, weights $p_i = 1$ in expression (4.8). Taking $\hat{G}$'s reversed roles of α and β into consideration, the `abc` call was

```
abc(TT, ahat, S, bhat, mu)
```
(6.28)

Table 6. abc algorithm calculations for the diabetes data, Fig. 3†

c	$\hat{\theta}$	$\widehat{sd}$	a	z_0	c_q	*abc results*		*abcq results*	
						lo	*up*	*lo*	*up*
0.04	0.00	0.01	0.00	0.27	1.50	0.00	0.06	0.00	0.03
0.08	0.01	0.03	0.01	0.21	1.17	0.00	0.13	0.01	0.09
0.12	0.04	0.08	0.00	0.12	0.88	0.00	0.25	0.02	0.22
0.16	0.11	0.19	0.00	0.05	0.60	0.02	0.44	0.03	0.45
0.2	0.25	0.32	0.00	0.02	0.33	0.05	0.63	0.04	0.68
0.24	0.46	0.40	0.00	−0.02	0.05	0.13	0.80	0.08	0.86
0.28	0.67	0.36	0.00	−0.02	−0.23	0.28	0.92	0.23	0.95
0.32	0.84	0.24	0.00	−0.03	−0.50	0.49	0.98	0.47	0.96
0.36	0.94	0.12	0.00	−0.03	−0.78	0.71	0.99	0.73	0.97

†(a, z_0, c_q) are the three coefficients that adjust the standard limits $\hat{\theta} \pm \widehat{\text{sd}}$ to second-order accuracy (DiCiccio and Efron, 1992). The `abc` limits (seventh and eighth columns) were not much different from the purely local `abcq` limits (ninth and 10th columns).

where `mu` was the function

$$\text{mu}(b) = \sum_{i=1}^{B} W_i(b)\alpha_i \Big/ \sum_{i=1}^{B} W_i(\beta) \qquad W_i(b) = \exp\{(b-\hat{\beta})^{\mathrm{T}}\alpha_i\}, \tag{6.29}$$

`bhat` $= \hat{\beta} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$, `ahat=mu(bhat)`, and `S` the $p \times p$ covariance matrix of the α_i; `TT` was the function

$$\text{TT}(a) = \sum_{i=1}^{B} W_i(b)t_{ci} \Big/ \sum_{i=1}^{B} W_i(b), \qquad b = \text{mu}^{-1}(a), \tag{6.30}$$

where $t_{ci} = t_c(\alpha_i)$ (2.31), and mu^{-1} was the inverse function of `mu`, calculated to accuracy 10^{-11} by using Newton–Raphson iteration. (The inversion is necessary because $\hat{\theta} = s(b)$ (4.12) is a function of the natural parameter b of $\hat{G}$, but `abc` requires $\hat{\theta}$ stated in terms of the expectation parameter: a in the case of $\hat{G}$.)

Table 6 displays a portion of the `abc` output going into Fig. 3. Besides $\hat{\theta}$ and $\widehat{\text{sd}}$, it shows the three second-order correction coefficients that were described in DiCiccio and Efron (1992): acceleration a and bias correction z_0 are mostly ignorable, but the quadratic coefficient c_q is not. It has a major effect on the `abcq` limits, which is a version of `abc` that is purely local in the sense of only recomputing $T(b)$ for b near $\hat{\beta}$.

The `abc` limits in Fig. 3 involve one non-local recomputation. They enjoy tranformation invariance, monotone transformations of the parameter of interest producing the same transformation of the interval end points, which might be helpful for parameters like θ_c restricted to the interval [0, 1]. However, in this case they were not much different from the `abcq` versions.

6.8. Remark 8: Tweedie's formula for the prostate data

Both Bayes and empirical Bayes hierarchical analyses require evaluation of $t_i = E\{\tau(\delta_0)|z_0, \alpha_i, \hat{\beta}\}$ (5.21) for $i = 1, 2, \ldots, B$. This is straightforward when $\tau(\delta) = \delta$ as in Fig. 6. *Tweedie's formula* (Efron, 2011) says that

$$E(\delta_0|z_0, \alpha) = z_0 + \frac{\mathrm{d}}{\mathrm{d}z}\log\{f_\alpha(z)\}|_{z_0}, \tag{6.31}$$

where $f_\alpha(z)$ is the marginal density (5.4). In terms of notation (5.11)–(5.12),

$$t_i = c_{j_0} + \dot{\mathbf{x}}_{j_0}\alpha_i, \tag{6.32}$$

where j_0 is the bin index (5.9) for z_0, and

$$\dot{\mathbf{x}}_j = (0, 1, 2c_j, 3c_j^2, \ldots, 8c_j^7). \tag{6.33}$$

Theoretically there is a version of Tweedie's formula applying to any function $\tau(\delta)$ (called 'Bayes rule in terms of f' in Efron (2014)). The case $\tau(\delta) = \delta$, however, is particularly favourable to generalized linear model modelling of the marginal density $f(z)$ (5.4). Other choices of $\tau(\delta)$ may require models for f that are not generalized linear models, returning hierarchical Bayes analysis to the general, non-exponential family framework of Section 2.

6.9. Remark 9: empirical Bayes standard deviation formula

The empirical Bayes standard deviation formula (5.27) is easy to derive in exponential families. We assume, for convenience, that the sufficient statistic x takes on only a finite number J of possible values, so that the marginal density $f_\alpha(\cdot)$ is represented by a J-vector $\mathbf{f}_\alpha$. Let $\mathbf{Q}$ be the gradient of $t(\alpha) = E\{\tau(\delta_0)|z_0, \alpha\}$ with respect to $\mathbf{f}$ (specific formulae for Q are given in Efron (2014)) and $\dot{\mathbf{f}}_\alpha$ the $J \times p$ derivative matrix $(\partial \mathbf{f}_{\alpha_j}/\partial\alpha_k)$. Then a first-order Taylor series expansion gives

$$t(\hat{\alpha}) - t(\alpha) \doteq \mathbf{Q}^{\mathrm{T}}\dot{\mathbf{f}}_\alpha(\hat{\alpha} - \alpha). \tag{6.34}$$

This yields

$$\mathrm{sd}\,\{t(\hat{\alpha})\}^2 \doteq \mathbf{Q}^{\mathrm{T}}\dot{\mathbf{f}}_\alpha\Sigma_\alpha\dot{\mathbf{f}}_\alpha\mathbf{Q} \qquad \Sigma_\alpha = \mathrm{cov}_\alpha(\hat{\alpha}) \tag{6.35}$$

and

$$\mathrm{cov}\,\{t(\hat{\alpha}), \hat{\alpha}\} \doteq \mathbf{Q}^{\mathrm{T}}\dot{\mathbf{f}}_\alpha\Sigma_\alpha, \tag{6.36}$$

so

$$\mathrm{sd}\{t(\hat{\alpha})\}^2 \doteq \mathrm{cov}\{t(\hat{\alpha}), \hat{\alpha}\}^{\mathrm{T}}\Sigma_\alpha^{-1}\,\mathrm{cov}\{t(\hat{\alpha}), \hat{\alpha}\}. \tag{6.37}$$

But $\Sigma_\alpha^{-1} \doteq \mathrm{cov}_\alpha(\hat{\beta}) = V_\alpha$ in exponential families, giving

$$\mathrm{sd}\{t(\hat{\alpha})\}^2 \doteq \mathrm{cov}\{t(\hat{\alpha}), \hat{\alpha}\}^{\mathrm{T}}V_\alpha\,\mathrm{cov}\{t(\hat{\alpha}), \hat{\alpha}\}. \tag{6.38}$$

Formula (5.27) for $\overline{\mathrm{sd}}$ is the bootstrap evaluation of expression (6.38).

Acknowledgements

This research was supported in part by National Institutes of Health grant 8R37 EB002784 and by National Science Foundation grant DMS 1208787.

References

Berger, J. (2006) The case for objective Bayesian analysis. *Baysn Anal.*, **1**, 385–402.
Berger, J. O. (1985) *Statistical Decision Theory and Bayesian Analysis*, 2nd edn. New York: Springer.
Carlin, B. P. and Louis, T. A. (2000) *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd edn. Boca Raton: Chapman and Hall–CRC.
Casella, G. and Berger, R. L. (2002) *Statistical Inference*, 2nd edn. Pacific Grove: Duxbury.
Diaconis, P. and Freedman, D. (1986) On the consistency of Bayes estimates (with discussion). *Ann. Statist.*, **14**, 1–67.
Diaconis, P. and Ylvisaker, D. (1979) Conjugate priors for exponential families. *Ann. Statist.*, **7**, 269–281.
DiCiccio, T. and Efron, B. (1992) More accurate confidence intervals in exponential families. *Biometrika*, **79**, 231–245.

Efron, B. (1987) Better bootstrap confidence intervals (with comments). *J. Am. Statist. Ass.*, **82**, 171–200.
Efron, B. (2011) Tweedie's formula and selection bias. *J. Am. Statist. Ass.*, **106**, 1602–1614.
Efron, B. (2012) Bayesian inference and the parametric bootstrap. *Ann. Appl. Statist.*, **6**, 1971–1997.
Efron, B. (2014) Two modeling strategies for empirical Bayes estimation. *Statist. Sci.*, to be published.
Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression (with discussion). *Ann. Statist.*, **32**, 407–499.
Fearnhead, P. and Prangle, D. (2012) Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation (with discussion). *J. R. Statist. Soc.* B, **74**, 419–474.
Fraser, D. A. S. (1990) Tail probabilities from observed likelihoods. *Biometrika*, **77**, 65–76.
Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995) *Bayesian Data Analysis*. New York: Chapman and Hall.
Ghosh, M. (2011) Objective priors: an introduction for frequentists (with discussion). *Statist. Sci.*, **26**, 187–202.
Johnstone, I. M. and Silverman, B. W. (2004) Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.*, **32**, 1594–1649.
Kass, R. E. and Wasserman, L. (1996) The selection of prior distributions by formal rules. *J. Am. Statist. Ass.*, **91**, 1343–1370.
Little, R. J. (2006) Calibrated Bayes: a Bayes/frequentist roadmap. *Am. Statistn*, **60**, 213–223.
Meneses, J., Antle, C. E., Bartholomew, M. J. and Lengerich, R. (1990) A simple algorithm for delta method variances for multinomial posterior Bayes probability estimates. *Communs Statist. Simuln Computn*, **19**, 837–845.
Morris, C. N. (1983) Parametric empirical Bayes inference: theory and applications (with discussion). *J. Am. Statist. Ass.*, **78**, 47–65.
Park, T. and Casella, G. (2008) The Bayesian lasso. *J. Am. Statist. Ass.*, **103**, 681–686.
Rice, J. A. (2007) *Mathematical Statistics and Data Analysis*, 3rd edn. Pacific Grove: Duxbury.
Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R. and Sellers, W. R. (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**, 203–209.
Spiegelhalter, D. J. and Smith, A. F. M. (1982) Bayes factors for linear and log-linear models with vague prior information. *J. R. Statist. Soc.* B, **44**, 377–387.
Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.
Welch, B. L. and Peers, H. W. (1963) On formulae for confidence points based on integrals of weighted likelihoods. *J. R. Statist. Soc.* B, **25**, 318–329.