FREQUENTIST ACCURACY OF BAYESIAN ESTIMATES

By

Bradley Efron

Technical Report 265
October 2013

Division of Biostatistics
STANFORD UNIVERSITY
Stanford, California

FREQUENTIST ACCURACY OF BAYESIAN ESTIMATES

By

Bradley Efron
Stanford University

Technical Report 265
October 2013

**This research was supported in part by
National Institutes of Health grant 8R37 EB002784.**

**Division of Biostatistics
STANFORD UNIVERSITY
Sequoia Hall, 390 Serra Mall
Stanford, CA 94305-4065**

http://statistics.stanford.edu

Frequentist accuracy of Bayesian estimates

Bradley Efron*†
Stanford University

Abstract

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. Bayes rule will still produce nice-looking estimates and credible intervals, but these lack the logical force attached to experience-based priors and require further justification. This 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 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: general accuracy formula, parametric bootstrap, abc intervals, hierarchical and empirical Bayes, MCMC

1 Introduction

The past two decades have witnessed the 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 $\theta = t(\mu)$, 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 θ given x,

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

How accurate is $\hat\theta$? The obvious answer, and the one 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, admissability, or esthetic Bayesian preference, but not trust what is after all a self-selected choice of prior as determining $\hat\theta$'s accuracy. Berger (2006) makes this point at the beginning of his Section 4.

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

*Research supported in part by NIH grant 8R37 EB002784 and by NSF grant DMS 1208787.
†Sequoia Hall, 390 Serra Mall, Stanford University, Stanford, CA 94305-4065; brad@stat.stanford.edu

Our main result, 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.

In fact several of our examples will demonstrate near equality between Bayesian and frequentist standard deviations. That does not have to be the case though; Remark 1 of 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 endpoint 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 connect Bayesian inference with the seemingly super-frequentist bootstrap world, 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, nicely reviewed in Chapter 4 of Carlin and Louis (2000). Particular attention focuses on large-sample behavior, where "the data swamps the prior" and $\hat{\theta}$ converges to the maximum likelihood estimator (see Result 8, 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 swamps the prior.

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).

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.

2 The 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 $\pi(\mu)$, while 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

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

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

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

(Proof to follow.)

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

$$\widehat{\mathrm{sd}} = \left[\operatorname{cov}\{t(\mu), \alpha_x(\mu)|x\}^\top V_{\hat{\mu}} \operatorname{cov}\{t(\mu), \alpha_x(\mu)|x\} \right]^{1/2}, \tag{2.6}$$

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

This is the *general accuracy formula*, general in the sense of applying to all choices of prior, uninformative or not. Section 3 shows that $\alpha_x(\mu)$ has a simple form, not depending on x, in exponential families.

The theorem is an immediate consequence of the lemma and the usual delta-method estimate for the standard deviation of a statistic $s(x)$: suppose for convenience that x is unbiased for μ, so that $m_\mu = E_\mu\{x\} = \mu$ and $\hat{\mu} = x$. A first-order Taylor series approximation yields

$$\mathrm{sd}_\mu\{s\} \doteq \left[s'(\mu)^\top V_\mu s'(\mu) \right]^{1/2}, \tag{2.7}$$

where $s'(\mu)$ indicates the gradient $\nabla_x s$ evaluated at $x = \mu$. Substituting $\hat{\mu}$ for μ gives the delta-method standard deviation estimate, for example in (2.6) where $s(x) = E\{t(\mu)|x\} = \hat{\theta}$. Posterior expectations tend to be smooth functions of x, even if $f_\mu(x)$ is not, improving the accuracy of the delta-method approximation. A useful special case of the theorem appeared in Meneses et al. (1990).

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

Implementation Suppose

$$\{\mu_1, \mu_2, \mu_3, \ldots, \mu_B\} \tag{2.8}$$

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),

$$t_i = t(\mu_i) \quad \text{and} \quad \alpha_i = \alpha_x(\mu_i). \tag{2.9}$$

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

$$\widehat{\mathrm{cov}} = \sum_{i=1}^{B} (\alpha_i - \bar{\alpha})(t_i - \bar{t})/B \qquad \left[\bar{\alpha} = \sum \alpha_i/B \right] \tag{2.10}$$

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.)

Posterior sampling The posterior sample $\{\mu_1, \mu_2, \ldots, \mu_B\}$ will typically be obtained via MCMC, after a suitable burn-in period. The nonindependence of the μ's does not invalidate (2.10), 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 (3.12).

Factorization If

$$f_\mu(x) = g_\mu(x)h(x) \tag{2.11}$$

then the gradient

$$\nabla_x \log f_\mu(x) = \nabla_x \log g_\mu(x) + \nabla_x \log h(x). \tag{2.12}$$

The last term does not depend on μ, so $\mathrm{cov}\{t(\mu), \nabla_x \log f_\mu(x)|x\}$ equals $\mathrm{cov}\{t(\mu), \nabla_x \log g_\mu(x)|x\}$ and we can take

$$\alpha_x(\mu) = \nabla_x \log g_\mu(x) \tag{2.13}$$

in the lemma and the theorem.

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

$$\alpha_x(\mu) = \nabla_x \log f_\mu(x) = \nabla_x \log g_\mu(y) + \nabla_x \log h(z). \tag{2.14}$$

As in (2.12), 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,

$$\alpha_x(\mu) = {Y'}^\top \alpha_y(\mu), \tag{2.15}$$

where Y' is the $q \times p$ matrix $(\partial y_i/\partial x_j)$. From (2.6) we get

$$\widehat{\mathrm{sd}} = \left[\mathrm{cov}\left\{t(\mu), \alpha_y(\mu)|y\right\}^\top Y' V_{\hat\mu} {Y'}^\top \mathrm{cov}\left\{t(\mu), \alpha_y(\mu)|y\right\}\right]^{1/2}. \tag{2.16}$$

Notice that $Y' V_{\hat\mu} {Y'}^\top$ is the delta-method estimate of the covariance matrix of y when μ equals $\hat\mu$. In this approximate sense the theorem 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 the case in the exponential family setup of Section 3.) More importantly, working with y makes $\widehat{\mathrm{cov}}$ in (2.10) lower-dimensional, and yields better estimation properties when substituted into (2.6).

Vector parameter of interest The lemma and theorem 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 $\mathrm{cov}\{t(\mu), \alpha_x(\mu)|x\}$ in (2.5) become $p \times K$ matrices, yielding $K \times K$ approximate frequentist covariance matrix $\widehat{\mathrm{var}}$ for $\hat\theta = E\{t(\mu)|x\}$,

$$\widehat{\mathrm{var}} = \mathrm{cov}\left\{t(\mu), \alpha_x(\mu)|x\right\}^\top V_{\hat\mu}\, \mathrm{cov}\left\{t(\mu), \alpha_x(\mu)|x\right\}, \tag{2.17}$$

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

Discrete statistic x Suppose $\mathcal{F}$ in (2.2) is the one-dimensional Poisson family $f_\mu(x) = \exp(-\mu)\cdot \mu^x/x!$, x a nonnegative integer. We can still calculate $\alpha_x(\mu) = \log(\mu)$ (2.4) (ignoring the term due to $x!$, as in (2.12)). For μ greater than, say, 10, the Poisson distribution ranges widely enough to smooth over its discrete nature, and we can expect formula (2.6) to apply reasonably well. Section 5 discusses a multidimensional discrete application.

Bias correction Replacing $\mathrm{cov}\{t(\mu), \alpha_x(\mu)|x\}$ in (2.6) with its nearly unbiased estimate $\widehat{\mathrm{cov}}$ (2.10) upwardly biases the sd 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 will consider the *Diabetes Data* of Efron et al. (2004): $n = 442$ diabetes patients each have had observed a vector $\boldsymbol{x}$ of $p = 10$ predictor variables (age, sex, body mass index, blood pressure, and six blood serum measurements),

$$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{i10}) \qquad \text{for } i = 1, 2, \ldots, n = 442, \tag{2.18}$$

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

$$\boldsymbol{y} = \boldsymbol{X}\alpha + \boldsymbol{e} \qquad \text{with } \boldsymbol{e} \sim \mathcal{N}_n(0, \boldsymbol{I}). \tag{2.19}$$

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

Park and Casella (2008) consider applying a Bayesian version of the lasso (Tibshirani, 1996) to the Diabetes Data. In terms of our model (2.19) (they do not standardize the variables) Park and Casella take the prior distribution for α to be

$$\pi(\alpha) = e^{-\lambda L_1(\alpha)}, \tag{2.20}$$

with $L_1(\alpha)$ the L_1 norm $\sum_1^{10} |\alpha_j|$, and λ having value

$$\lambda = 0.37. \tag{2.21}$$

The Laplace-type prior (2.20) results in the posterior mode of β given $\boldsymbol{y}$ coinciding with the lasso estimate

$$\hat{\alpha}_\lambda = \arg\min_\alpha \left\{ \|\boldsymbol{y} - \boldsymbol{X}\alpha\|^2/2 + \lambda L_1(\alpha) \right\}, \tag{2.22}$$

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 will ignore that here and assume prior (2.20)–(2.21) to be pre-selected.

An MCMC algorithm was used to produce (after burn-in) $B = 10,000$ samples α_i from the posterior distribution $\pi(\alpha|\boldsymbol{y})$, under assumptions (2.19)–(2.21),

$$\{\alpha_i,\ i = 1, 2, \ldots, B = 10,000\}. \tag{2.23}$$

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

$$\hat{\theta} = \sum_{i=1}^{B} t(\alpha_i)/B. \tag{2.24}$$

We can then apply Theorem 1 to estimate the frequentist standard deviation of $\hat{\theta}$. In terms of the general notation (2.2), α is the parameter μ and we can take the sufficient statistic $\hat{\beta} = \boldsymbol{X}'\boldsymbol{y}$ in model (2.19) to be x. With $\widehat{\text{cov}}$ defined as in (2.10) formula (2.6), using the MCMC sample (2.23), gives

$$\widehat{\text{sd}} = \left\{\widehat{\text{cov}}^{\top} G\, \widehat{\text{cov}}\right\}^{1/2} \qquad \left[G = \boldsymbol{X}^{\top}\boldsymbol{X}\right], \tag{2.25}$$

since $G = \boldsymbol{X}^{\top}\boldsymbol{X}$ is the covariance matrix of $\hat{\beta}$. The fact that α in (2.19) can substitute for $\alpha_x(\mu)$ in (2.6), easily verified from $\hat{\beta}|\alpha \sim \mathcal{N}(G\alpha, G)$, is an example of the general exponential family result in Theorem 2 of Section 3.

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

$$\gamma_{125} = \boldsymbol{x}_{125}\alpha, \tag{2.26}$$

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

$$\mathcal{N}(0.248, 0.072^2). \tag{2.27}$$

Formula (2.25) gave frequentist standard deviation 0.071 for $\hat{\theta}_{125} = 0.248 = \sum \hat{\gamma}_{125,i}/10,000$, almost the same as the posterior standard deviation, but having a quite different interpretation. The near equality here is no fluke, but can turn out differently for other linear combinations $\gamma = \boldsymbol{x}\alpha$; see Remark 1 of Section 6. *Note*: It is helpful here and in what follows to denote the parameter of interest as γ with $\hat{\theta} = E\{\gamma|x\}$ indicating its posterior expectation.

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

$$t_c(\alpha) = \begin{cases} 1 & \text{if } \boldsymbol{x}_{125}\alpha \le c \\ 0 & \text{if } \boldsymbol{x}_{125}\alpha > c \end{cases} \tag{2.28}$$

so $E\{t_c(\alpha)|\boldsymbol{y}\} = \Pr\{\gamma_{125} \le c|\boldsymbol{y}\}$. The MCMC sample (2.23) provides $B = 10,000$ posterior values t_{ci}, from which we obtain the estimated cdf(c) value $\sum_1^B t_{ci}/B$ and its standard deviation (2.6); for example $c = 0.3$ gives

$$\Pr\{\gamma_{125} \le 0.3|\boldsymbol{y}\} = 0.762 \pm 0.304, \tag{2.29}$$

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

The heavy curve in Figure 1 traces the posterior cdf of γ_{125}. Dashed vertical bars indicate $\pm$ one frequentist standard deviation. If we take the prior (2.20) literally then the cdf curve is exact, but if not, the large frequentist standard errors suggest cautious interpretation.

The cdf curve equals 0.90 at $\hat{c} = 0.342$, this being the upper endpoint of a one-sided Bayes 90% credible interval. The frequentist standard deviation of $\hat{c}$ is 0.069 (obtained from $\widehat{\text{sd}}(\text{cdf}(\hat{c}))$ 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 were able to compare the frequentist standard deviation 0.071 with its Bayes posterior counterpart 0.072 (2.27). No such comparison is possible for the posterior cdf estimates: the cdf curve in Figure 1 is exact under prior (2.20)–(2.21). We might add a hierarchical layer of Bayesian assumptions in front of (2.20)–(2.21) in order to assess the curve's variability, but it is not obvious how to do so here. (Park and Casella, 2008, Section 3.2, make one suggestion.)

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

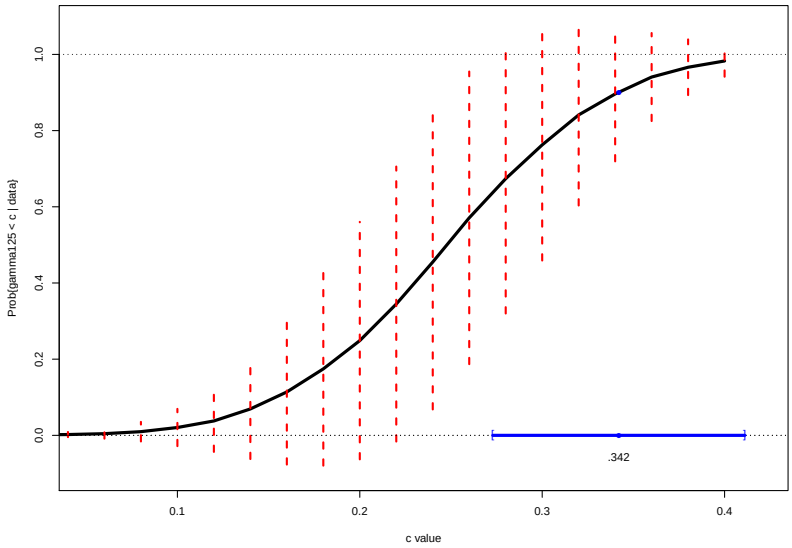

Figure 1: Heavy curve is posterior cdf of γ_{125} (2.26), Diabetes Data; vertical dashed lines indicate $\pm$ one frequentist standard error. The estimated curve is quite uncertain from a frequentist viewpoint. The upper 0.90 value $\hat{c} = 0.342$ has frequentist standard error 0.069, as indicated by the horizontal bar.

Proof of Lemma 1. Write $\hat{\theta} = A(x)/B(x)$ where

$$A(x) = \int_\Omega t(\mu)\pi(\mu)f_\mu(x)\ d\mu \quad \text{and} \quad B(x) = \int_\Omega \pi(\mu)f_\mu(x)\ d\mu. \tag{2.30}$$

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

$$\begin{aligned} A'(x) &= \int_\Omega t(\mu)\alpha_x(\mu)\pi(\mu)f_\mu(x)\ d\mu, \\ B'(x) &= \int \alpha_x(\mu)\pi(\mu)f_\mu(x)\ d\mu. \end{aligned} \tag{2.31}$$

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

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

■

An overly generous sufficient condition for the interchange of integration and differentiation in (2.31) is that $\alpha_x(\mu)$ and $t(\mu)$ are bounded in an open neighborhood of x and in a set of $\pi(\mu)$ probability 1. See Remark B of Section 6 for a more computational derivation of Lemma 1. An equivalent bootstrap form of the lemma and theorem is the subject of the next section.

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 $\pi(\mu|x)$ (2.8). 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 connect Bayesian and frequentist points of view.

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

$$\mathcal{F} : \left\{ f_\alpha\left(\hat\beta\right) = e^{\alpha^\top \hat\beta - \psi(\alpha)} f_0\left(\hat\beta\right),\ \alpha \in \mathcal{A} \right\}. \tag{3.1}$$

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 $\hat\beta$ being the maximum likelihood estimate (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 general 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 (2.6) takes a simplified form in exponential families, where the gradient (2.4) becomes

$$\begin{aligned} \nabla_{\hat\beta} \log f_\alpha\left(\hat\beta\right) &= \nabla_{\hat\beta}\left\{ \alpha^\top \hat\beta - \psi(\alpha) + \log f_0\left(\hat\beta\right)\right\} \\ &= \alpha + \nabla_{\hat\beta} \log f_0\left(\hat\beta\right). \end{aligned} \tag{3.3}$$

The final term does not depend on α so, as in (2.12), what was called $\alpha_x(\mu)$ in (2.4) is now simply α. Given prior distribution $\pi(\alpha)$ and parameter of interest $t(\alpha)$, we get this version of Theorem 1.

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*

$$\widehat{\mathrm{sd}} = \left[\mathrm{cov}\left\{ t(\alpha), \alpha|\hat\beta\right\}^\top V_{\hat\alpha}\, \mathrm{cov}\left\{ t(\alpha), \alpha|\hat\beta\right\}\right]^{1/2}, \tag{3.4}$$

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

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 $\hat\alpha$,

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

(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"

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

the ratio of the likelihood to the bootstrap density. (See (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 $\hat{\theta} = \sum_1^B t_i p_i$, so

$$\hat{\theta} = \left[\sum_{i=1}^{B} t_i \pi_i f_{\alpha_i}\left(\hat{\beta}\right) \Big/ f_{\hat{\alpha}}(\beta_i)\right] \Big/ \left[\sum_{i=1}^{B} \pi_i f_{\alpha_i}\left(\hat{\beta}\right) \Big/ f_{\hat{\alpha}}(\beta_i)\right]. \tag{3.8}$$

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

$$\left[\int_{\mathcal{A}} t(\alpha)\pi(\alpha) f_\alpha\left(\hat{\beta}\right)\, d\alpha\right] \Big/ \left[\int_{\mathcal{A}} \pi(\alpha) f_\alpha\left(\hat{\beta}\right)\right], \tag{3.9}$$

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

The same argument applies to any posterior calculation. In particular, $\operatorname{cov}\{t(\alpha), \alpha|\hat{\beta}\}$ in (3.4) is approximated by

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

Implementing Theorem 2 now follows three automatic steps:

1. Generate a parametric bootstrap sample $\beta_1, \beta_2, \ldots, \beta_B$ (3.5).
2. For each β_i calculate $\alpha_i,\ t_i = t(\alpha_i)$, and p_i (3.7).
3. Compute $\widehat{\operatorname{cov}}$ (3.10).

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

$$\widehat{\mathrm{sd}} = \left[\widehat{\operatorname{cov}}^\top V_{\hat{\alpha}}\, \widehat{\operatorname{cov}}\right]^{1/2}. \tag{3.11}$$

(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 $\widehat{\mathrm{sd}}$.

An MCMC implementation sample $\{\mu_i,\ i = 1, 2, \ldots, B\}$ (2.8) approximates a multidimensional posterior distribution by an equally weighted distribution on B nonindependent points. The bootstrap implementation (3.5)–(3.7) puts *unequal* weights on B i.i.d. (independent and identically distributed) 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 calculations show that $\hat{\theta} = \sum Q_i / \sum P_i$ has internal squared coefficient of variation approximately

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

$\bar{Q} = \sum Q_i/B$ and $\bar{P} = \sum P_i/B$. See Remark 3 of Section 6.

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:

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

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

$$\Delta(\alpha) = (\alpha - \hat{\alpha})^\top \left(\beta + \hat{\beta}\right) - 2\left[\psi(\alpha) - \psi\left(\hat{\alpha}\right)\right], \tag{3.14}$$

and, to a good approximation,

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

with $\pi^{\text{Jeff}}(\alpha) = |V_\alpha|^{1/2}$, Jeffreys invariant prior for α (Lemma 1, Efron, 2012). If our prior $\pi(\alpha)$ *is* $\pi^{\text{Jeff}}(\alpha)$ then

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

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 one 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 prior (2.20)–(2.21) for the Diabetes Data would be a poor choice for bootstrap implementation (though this difficulty can be mitigated by recentering the parametric bootstrap resampling distribution).

Table 1: *Cell infusion data.* Human cell colonies were infused with mouse nucleii in 5 different proportions, over time periods varying from 1 to 5 days, and observed to see if they did or did not thrive. The table displays number thriving over number of colonies. For example, 5 of the 31 colonies in the lowest infusion/days category thrived.

	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

Table 1 displays the *cell infusion data*, which we will use to illustrate bootstrap implementation of the general accuracy formula. Human muscle cell colonies were infused with mouse nucleii. Five increasing infusion proportions of mouse nucleii were tried, cultured over time periods ranging from one to five days, and observed to see if they thrived or did not. The table shows, for instance, that 52 of the 61 colonies in the highest proportion/days category thrived.

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 fit 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 will take

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

the ratio of overall probability of success on Day 5 compared to Day 1, and calculate its posterior distribution assuming Jeffreys prior $\pi^{\text{Jeff}}(\alpha)$ on α.

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

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

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 (3.19). The weights p_i (3.7) that convert the bootstrap sample into a posterior distribution are

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

according to (3.7) and (3.16). Here Δ_i is the half binomial deviance difference; see Remark 5, Section 6.

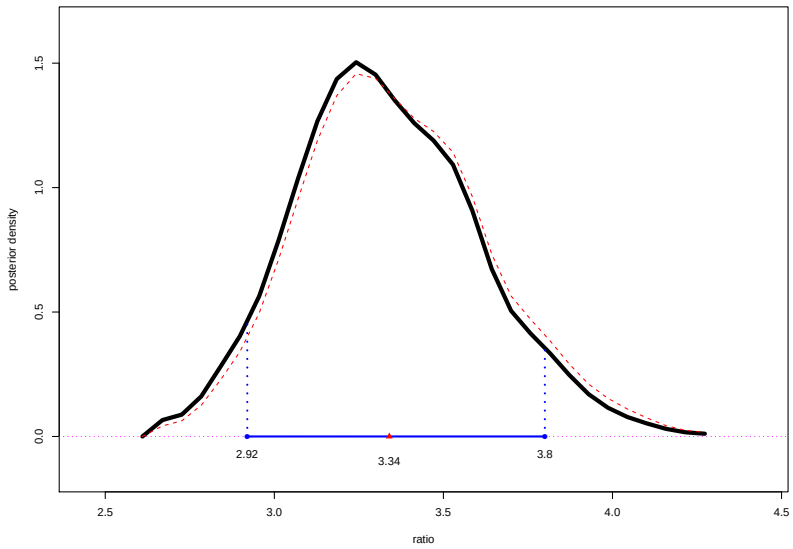

Figure 2: Posterior density of ratio γ (3.19) given the cell infusion data; binomial model (3.17)–(3.18), Jeffries prior $\pi^{\text{Jeff}}(\alpha)$. From $B = 2000$ parametric bootstrap replications (3.20), posterior expectation 3.34 has frequentist $\widehat{\text{sd}} = 0.273$. Solid line segment shows central 0.90 credible interval $[2.92, 3.80]$. Frequentist sd of 0.90 content is 0.042. Light dashed line is unweighted bootstrap density.

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

$$\hat{\theta} = \sum_{1}^{B} p_i t(\alpha_i) = 3.34 \tag{3.22}$$

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

How accurate is $\hat{\theta}$? Formula (3.11) yields $\widehat{\text{sd}} = 0.273$ as its frequentist standard deviation. This is almost the same as the Bayes posterior standard deviation $[\sum p_i(\gamma_i - \hat{\theta})^2]^{1/2} = 0.266$.

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, 2013). The bootstrap estimate of standard deviation $[\sum(\gamma_i - \bar{\gamma})^2]^{1/2}$ equals 0.270, and it is not surprising that both the Bayes posterior sd and the frequentist delta-method sd are close to 0.270.

Integrating the posterior density curve in Figure 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 $\widehat{\text{sd}} = 0.042$ for the frequentist standard deviation of the interval's content. The two endpoints have sd's 0.22 and 0.31. More interestingly, their frequentist correlation (calculated using (2.17); 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.

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 Figure 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

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

in the exponential family setup (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 BC_a interval (Efron, 1987). By construction, such intervals would *not* protrude beyond $[0, 1]$ in the equivalent of Figure 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 on the order of 1000. Each of these would require further simulations, $\{\mu_1, \mu_2, \ldots, \mu_B\}$ as in (2.8) or $\{\beta_1, \beta_2, \ldots, \beta_B\}$ as in (3.5), B also exceeding 1000, in order to accurately calculate the $\hat{\theta}_k^*$.

A shortcut method for bootstrap confidence calculations that, like Theorems 1 and 2, requires *no* additional replications, will be developed next. The shortcut requires exponential family structure (3.1), but otherwise applies equally to MCMC or bootstrap implementation (2.8) or (3.5).

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

$$g\left(\alpha|\hat{\beta}\right) = \pi(\alpha) f_\alpha\left(\hat{\beta}\right) \Big/ f\left(\hat{\beta}\right) \qquad \left[f\left(\hat{\beta}\right) = \int_{\mathcal{A}} \pi(\alpha) f_\alpha\left(\hat{\beta}\right)\ m(d\alpha)\right] \tag{4.2}$$

(with $m(\cdot)$ the underlying measure for the family $\mathcal{F}$, often Lebesgue measure or counting measure

on a discrete set). Suppose 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 $\mathcal{G}$,*

$$\mathcal{G} = \left\{ g(\alpha|b) = e^{\left(b-\hat{\beta}\right)^\top \alpha - \phi(b)} g\left(\alpha|\hat{\beta}\right) \text{ for } b - \hat{\beta} \in \hat{\mathcal{B}} \right\}, \tag{4.3}$$

where

$$e^{\phi(b)} = \int_{\mathcal{A}} e^{\left(b-\hat{\beta}\right)^\top \alpha} g\left(\alpha|\hat{\beta}\right)\ m(d\alpha). \tag{4.4}$$

$\mathcal{G}$ 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 (4.4) *is finite.*

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

Proof. From (3.1) we compute

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

But

$$f_\alpha(b)/f_\alpha\left(\hat{\beta}\right) = e^{\left(b-\hat{\beta}\right)^\top \alpha} \left[f_0(b) \Big/ f_0\left(\hat{\beta}\right)\right], \tag{4.6}$$

yielding

$$g(\alpha|b) = g\left(\alpha|\hat{\beta}\right) e^{\left(b-\hat{\beta}\right)^\top \alpha} \left[\frac{f_0(b) f\left(\hat{\beta}\right)}{f_0\left(\hat{\beta}\right) f(b)}\right]. \tag{4.7}$$

The final factor does not depend on α and so must equal $\exp(-\phi(b))$ in (4.3)–(4.4) in order for (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

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

$p_i = \pi_i R_i / \sum_1^B \pi_j R_j$ in bootstrap implementation (3.5)–(3.9); and $p_i = 1/B$ in the MCMC implementation (2.8) where the μ_i play the role of the α_i.

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

$$W_i(b) = e^{\left(b-\hat{\beta}\right)^\top \alpha_i}. \tag{4.9}$$

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: $\hat{\mathcal{G}}$ differs from the empirical exponential family in Section 6 of Efron, 2013.)

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 $\hat{g}(\alpha_i|\hat{b}^*_k)$ as

$$\hat{\theta}^*_k = \sum_{i=1}^{B} t_i W_i\left(\hat{b}^*_k\right) p_i \Big/ \sum_{i=1}^{B} W_i\left(\hat{b}^*_k\right) p_i \qquad [t_i = t(\alpha_i)]\,. \tag{4.11}$$

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 $\hat{\theta} = \sum t_i p_i$, 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 (2013) 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*, that is, by considering alternate data values b very close to the actual $\hat{\beta}$, rather than full bootstrap resamples as in (4.1).

This brings us to the *abc* system of confidence intervals ("approximate bootstrap confidence," DiCiccio and Efron, 1992, not to be confused with "Approximate Bayesian Computation," as in Fearnhead and Prangle, 2012). The abc algorithm approximates full bootstrap confidence intervals using only a small number of resamples b in the immediate neighborhood of the observed sufficient statistic $\hat{\beta}$.

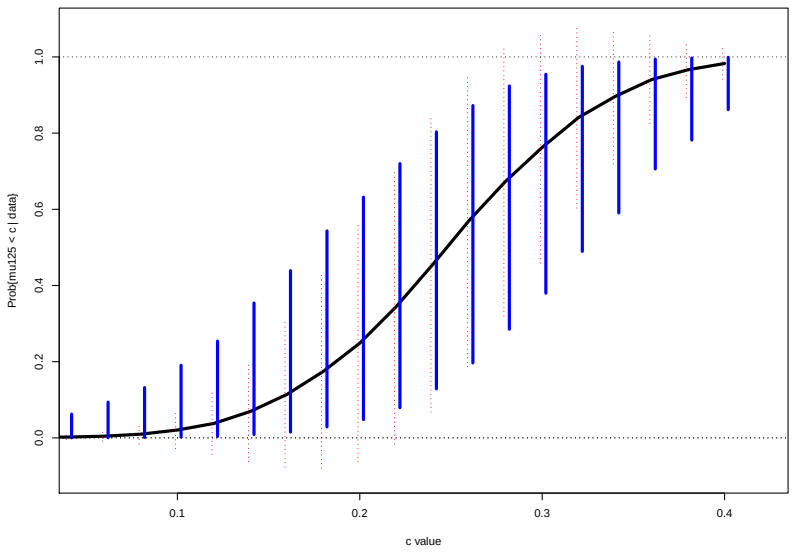

Figure 3: Vertical bars are 68% central abc confidence limits for patient 125's posterior cdf, Figure 1. They remain within the feasible interval $[0, 1]$, unlike Figure 1's standard deviation bars, shown here as light dashed vertical lines.

Figure 3 shows again the posterior cdf from Figure 1 for γ_{125}, the progression parameter for patient 125 in the diabetes study. The heavy vertical bars indicate abc 68% central frequentist confidence limits for the Bayes posterior cdf values. Now the confidence limits stay within $[0, 1]$. Remark 6 of Section 6 discusses the details of the abc calculations.

Standard confidence intervals, say $\hat{\theta} \pm \widehat{\text{sd}}$ for approximate 68% coverage, require only the original point estimate $\hat{\theta}$ and its accuracy estimate $\widehat{\text{sd}}$, 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 sample size n grows large.

The abc algorithm provides *second order accuracy*, that is, coverage errors approaching zero at rate n^{-1}. This is more than a theoretical nicety. As the examples in DiCiccio and Efron (1992) show, 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 endpoints), and for nonnormality. The algorithm depends on exponential family structure, provided by $\hat{\mathcal{G}}$ the empirical posterior family (4.10), and a smoothly varying point estimate.

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 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*,

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

depends on the prior $\pi(\alpha)$. It seems reasonable to ask that an uninformative prior not produce large definitional biases. The parameter γ_{125} (2.26) has MLE and standard deviation 0.316 ± 0.076, compared with its Bayes estimate and frequentist standard deviation 0.248 ± 0.071, giving a relative difference of

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

In other words, the Park and Casella prior (2.20) shifts the estimate for patient 125 about 0.9 standard deviations downward, a quite substantial effect.

Figure 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 (2.20) looks like a strong shrinker, and not at all uninformative, its overall effects on the patient estimates are 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 analyzing large parallel data sets. This section compares the accuracy of the two methods, providing some intuition as to 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

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

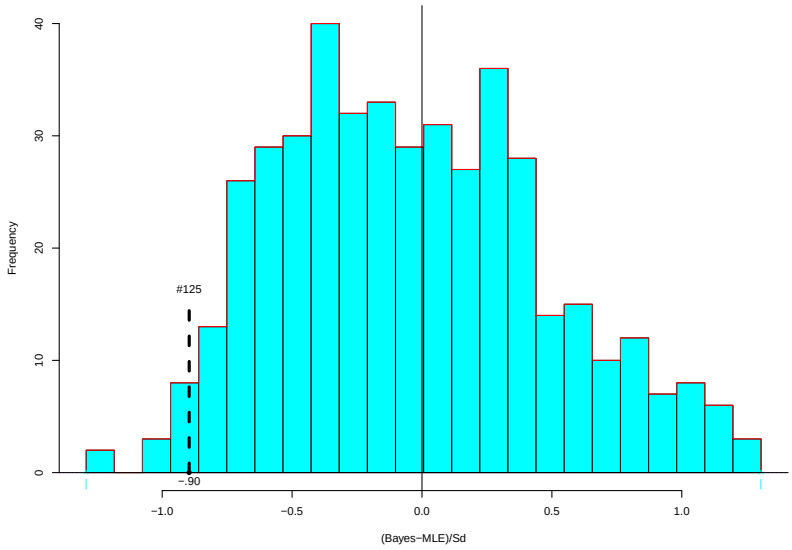

Figure 4: Relative differences (4.14) for the 442 diabetes patients, Park and Casella prior (2.20): Bayes estimate minus MLE, divided by MLE standard deviation.

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

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

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 will 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 $\boldsymbol{z}$.

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,

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

The empirical Bayes approach estimates the unknown value of α from the observed vector $\boldsymbol{z}$, 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_k's 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 $\boldsymbol{z}$,

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

We will employ the general accuracy formula to compare the frequentist variability of the two approaches.

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 patients with controls has been calculated for each gene, which we will 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 (5.3)–(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 nonnull genes, those having large δ_k values, either positive or negative.

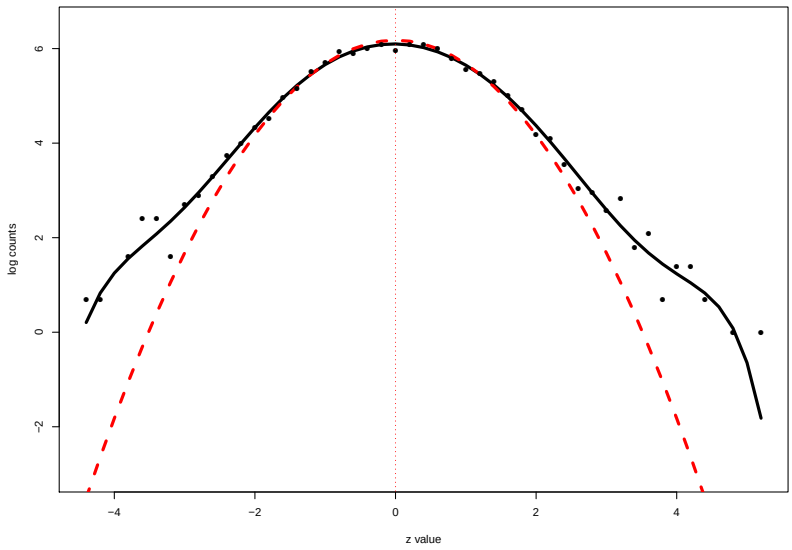

Figure 5: Prostate data: dots are log counts for 49 bins (5.8)–(5.9). Dashed quadratic curve would fit the dots if all genes were null, $\delta_k = 0$ in (5.7). Eighth-degree polynomial, heavy curve, gives a much better fit, indicating that some genes have large effect sizes.

Binning the data simplifies the Bayes and empirical Bayes analyses. For Figure 5 the data has been put into $J = 49$ bins $\mathcal{Z}_j$, each of width 0.2, with centers 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 Figure 5 are the log counts $\log(y_j)$. The dashed 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 solid curve, provided a good fit to the data. It was obtained from a Poisson regression GLM (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

$$\mu_j = E_\alpha\{y_j\} = e^{\boldsymbol{x}(c_j)\alpha}. \tag{5.11}$$

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

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

the c_j being the bin centers (5.8), while 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 $\boldsymbol{z}$ to the binned counts that we will 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

$$\hat\beta = \boldsymbol{X}^\top \boldsymbol{y}, \tag{5.13}$$

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

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

$\operatorname{diag}(\boldsymbol{\mu}_\alpha)$ 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 (3.4). A crucial feature of our hierarchical model 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 (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 $\hat\beta$,

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

In the prostate 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

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

as compared to the empirical Bayes MLE estimate $t(\hat\alpha)$.

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, obtained in the prostate study by Poisson regression from model (5.10)–(5.12), `glm(`$\boldsymbol{y} \sim \boldsymbol{X}$`, poisson)$coef` in language R. From $\hat\alpha$ we obtain parametric bootstrap samples $\boldsymbol{y}_i^*$, $i = 1, 2, \ldots, B$,

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

where $\hat\mu_j = \exp(\boldsymbol{x}(c_j)\hat\alpha)$. The $\boldsymbol{y}_i^*$ vector yields β_i and α_i, (3.5) and (3.6): $\beta_i = \boldsymbol{X}^\top \boldsymbol{y}^*$ and $\alpha_i =$ `glm(`$\boldsymbol{y}^* \sim \boldsymbol{X}$`, poisson)$coef`.

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

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

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

$$\Delta_i = (\alpha_i - \hat\alpha)^\top \left(\beta_i + \hat\beta\right) - 2\sum_{j=1}^{J}(\mu_{ij} - \hat\mu_j), \tag{5.20}$$

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

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

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

$$\hat{\theta} = \sum_{i=1}^{B} p_i t_i \qquad \left[p_i = e^{\Delta_i} \Big/ \sum_{k=1}^{B} e^{\Delta_k}\right], \tag{5.22}$$

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

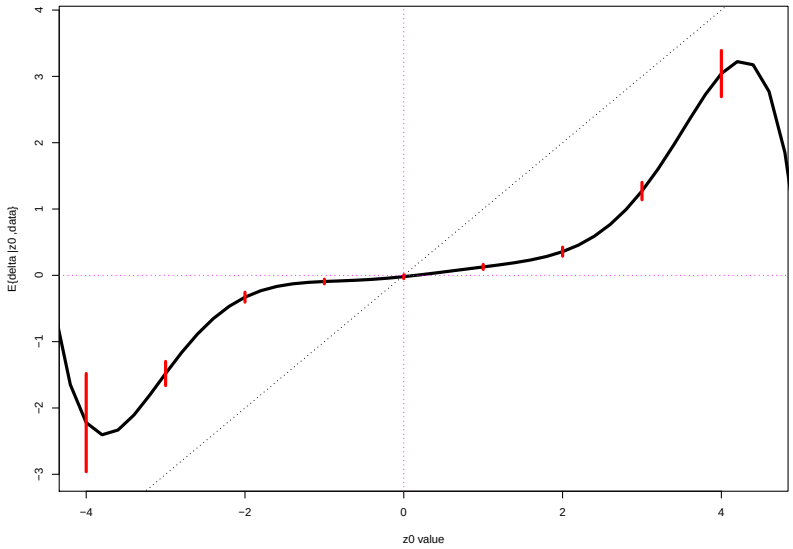

Figure 6: Hierarchical Bayes estimate $\hat{\theta} = E\{\delta_0|z_0, \hat{\beta}\}$ as a function of z_0, prostate study data. Vertical bars indicate $\pm$ one frequentist standard deviation $\widehat{\text{sd}}$ (3.11). Calculated from $B = 4000$ parametric bootstrap samples (5.18).

Figure 6 applies to the prostate data, taking $\tau(\delta)$, the function of interest, to be δ itself; that is, the hierarchical Bayes estimate (5.17) is

$$\hat{\theta} = E\left\{\delta_0|z_0, \hat{\beta}\right\}, \tag{5.23}$$

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 (5.19)–(5.22).

The heavy curve in Figure 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 zero, indicating nonnull effect sizes for large $|z_0|$, but always with strong "regression to the mean" behavior: $|\hat{\theta}| < |z_0|$. The vertical bars span plus or minus one frequentist standard deviation $\widehat{\text{sd}}$ (3.11).

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

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

Table 2: Comparison of hierarchical and empirical Bayes estimates for expected effect sizes in the prostate study. (1) Bayes estimate $\hat{\theta}$ (5.23); (2) empirical Bayes estimate $E\{\delta_0|z_0, \alpha = \hat{\alpha}, \hat{\beta}\}$; (3) Bayes posterior sd $[\sum p_i(t_i - \hat{\theta})^2]^{1/2}$; (4) frequentist sd of $\hat{\theta}$ (3.11); (5) bootstrap sd (5.25).

	z_0	-4	-3	-2	-1	0	1	2	3	4
1.	Bayes est	-2.221	-1.480	$-.329$	$-.093$	$-.020$	.127	.357	1.271	3.042
2.	Emp Bayes est	-2.217	-1.478	$-.331$	$-.092$	$-.020$	.126	.360	1.266	3.021
3.	Bayes sd	.756	.183	.074	.036	.030	.039	.071	.131	.336
4.	Bayes freq sd	.740	.183	.075	.035	.029	.038	.068	.131	.349
5.	Emp Bayes sd	.878	.187	.074	.037	.030	.039	.072	.139	.386

follows the curve in Figure 6 to within the line width. Table 2 gives numerical comparisons for $z_0 = -4, -3, \ldots, 4$. The estimated standard deviations for the empirical Bayes estimates, line 5, are a little bigger than those on line 3 for hierarchical Bayes, but that may just reflect the fact that the former are full bootstrap estimates while the latter are delta-method sd's.

Particularly striking is the agreement between the frequentist sd estimates for $\hat{\theta}$ (3.11), line 4, and the posterior Bayes sd estimates, line 3. This is the predicted *asymptotic* behavior (Berger, 1985, Sect. 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 then (5.23).

The empirical Bayes estimate $t(\hat{\alpha}) = E\{\delta_0|z_0, \alpha = \hat{\alpha}, \hat{\beta}\}$ had its standard deviation $\overline{\text{sd}}$, line 5 of Table 2, calculated directly from its bootstrap replications,

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

as compared with the Bayes posterior standard deviation, line 3,

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

(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, small in Table 2, depends on the empirical correlation between p_i and t_i.

There is a similar relation between lines 4 and 5 of the table. Remark 9 shows that $\overline{\text{sd}}$, line 5, is approximated by

$$\overline{\text{sd}} \doteq \left[\overline{\text{cov}}^\top V_{\hat{\alpha}} \overline{\text{cov}}\right]^{1/2}, \tag{5.27}$$

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

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

This compares with the weighted version (3.10)–(3.11) of line 4. Weighting did not matter much in Table 2, leaving the three standard deviations more alike than different.

Table 3: Polynomial model selection for the prostate study data. *Row 1*: raw bootstrap proportions for best polynomial fit, AIC criterion; *row 2*: corresponding Bayes posterior probabilities, Jeffreys prior; *row 3*: frequentist standard deviations for the Bayes estimates.

	m	4	5	6	7	8
1.	Bootstrap %	32%	10%	5%	1%	51%
2.	Bayes exp	36%	12%	5%	2%	45%
3.	Freq sd	±32%	±16%	±8%	±3%	±40%

The eighth-degree polynomial fit used in Figure 5 might be excessive. For each of the $B = 4000$ bootstrap samples $\boldsymbol{y}_i^*$, the "best" polynomial degree m_i^* was selected according to the AIC criterion, as detailed in Section 5 of Efron (2012). Only degrees $m = 0$ through 8 were considered. The top row of Table 3 shows that 32% of the 4000 bootstrap samples gave $m_i^* = 4$, compared to 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% on row 2.

We can apply Theorem 2 (3.11) to obtain frequentist standard deviations for the $\hat{\theta}^{(m)}$. These are shown in row 3. 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 (2.17)). 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 considered above, is frequentistically unstable for the prostate data.

6 Remarks

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

Remark 1. *Relation of Bayes and frequentist standard deviations* In several of our examples the posterior Bayes estimate $\hat{\theta}$ had its posterior standard deviation $\widehat{\text{sd}}_{\text{Bayes}}$ quite close to $\widehat{\text{sd}}_{\text{freq}}$, the frequentist sd. Why this might happen, or might not, is easy to understand in the diabetes data example (2.26)–(2.27).

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

$$\Sigma_\alpha = \tilde{\boldsymbol{\alpha}}^\top \tilde{\boldsymbol{\alpha}}/B \qquad (B = 10,000) \tag{6.1}$$

is the empirical covariance matrix of the α_i vectors. For any fixed row vector $\boldsymbol{x}$ we define as our parameter of special interest $\gamma_{\boldsymbol{x}} = \boldsymbol{x}\alpha$ ($\boldsymbol{x} = \boldsymbol{x}_{125}$ in (2.26)). Each α_i gives $t_i = \boldsymbol{x}\alpha_i$, with average $\bar{t} = \boldsymbol{x}\bar{\alpha}$. The vector $\tilde{\boldsymbol{t}}$ of centered values $\tilde{t}_i = t_i - \bar{t}$ is given by

$$\tilde{\boldsymbol{t}} = \tilde{\boldsymbol{\alpha}}\boldsymbol{x}^\top. \tag{6.2}$$

Then

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

Also, from (2.10),

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

yielding

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

from (2.25).

The variance ratio $\mathrm{rat}(\boldsymbol{x})$ equals

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

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

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

Table 4: Eigenvalue d_i for the variance ratio $\mathrm{rat}(\boldsymbol{x})$ (6.7).

1.014	1.009	.986	.976	.961	.944	.822	.710	.482	.098

Table 4 shows the eigenvalues d_i. We see that $\mathrm{rat}(\boldsymbol{x})$ could vary from 1.014 down to 0.098. For the 442 diabetes patients, $\mathrm{rat}(\boldsymbol{x}_i)$ ranged from 0.991 to 0.670, averaging 0.903; $\mathrm{rat}(\boldsymbol{x}_{125}) = 0.962$ was near the high end. A spherically uniform choice of $\boldsymbol{v}$ in (6.7) would yield an average $\mathrm{rat}(\boldsymbol{x})$ of 0.800.

The fact that the eigenvalues in Table 4 are mostly less than one relates to the Park and Casella prior (2.20). A flat prior for model (2.19) has $\mathrm{cov}(\boldsymbol{\alpha}) = G^{-1}$, giving $H = I$ and eigenvalues $d_i = 1$ in (6.7). The Park and Casella prior (2.20) is a "shrinker," making Σ_α and H less than I.

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 equaling $1/B$ or (3.7) in Sections 2 and 3, respectively, giving $\bar{t} = \sum p_i t_i$ and $\bar{\alpha} = \sum p_i\alpha_i$. Define $s_i = \sqrt{p_i}(t_i - \bar{t})$ and matrix M,

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

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

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

for $\hat\theta = \sum p_i t_i$; 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.

Remark 2. *A computational verification of Lemma 1* Working directly with the implementation values μ_i, α_i, and t_i (2.8)–(2.9), 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

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

x the observed statistic. Letting $\tilde x = x + dx$ with $dx \to 0$,

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

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

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

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

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

$$\hat\theta(x + dx) = \frac{\bar t\left[1 + \left(\overline{t\alpha}/\bar t\right)dx + \overline{tr}/\bar t\right]}{1 + \bar\alpha dx + \bar r}. \tag{6.13}$$

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

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

$\widehat{\mathrm{cov}}$ as in (2.10). This verifies Lemma 1 as employed in the computational form of Theorem 1: $\widehat{\mathrm{sd}} = (\widehat{\mathrm{cov}}^\top V_{\hat\mu}\widehat{\mathrm{cov}})^{1/2}$ (3.11).

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

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

where cov_* is the usual *unweighted* bootstrap covariance

$$\mathrm{cov}_* = \sum_{i=1}^B (\alpha_i - \bar\alpha)\,O_i/B \qquad \left[\bar\alpha = \sum_1^B \alpha_i/B\right]. \tag{6.17}$$

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

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

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

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

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

Then formula (6.18) can be reexpressed 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 (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 infinity. Combining (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 sd estimate when $B \to \infty$), indicating an upward bias in $\widehat{\mathrm{sd}}$.

The bias-corrected sd 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 (6.19) but with the jth group of replications removed, giving the $J \times p$ matrix $\boldsymbol{C}$ with rows C_{Bj}; finally the $p \times p$ sample covariance matrix of $\boldsymbol{C}$ gives the estimate

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

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

Remark 5. *Binomial deviance difference* The binomial GLM for the cell infusion data analysis (3.17)–(3.18), has half deviance difference

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

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

Remark 6. *A vector parameter example* The joint frequentist behavior of the 0.90 credible interval endpoints $[0.292, 0.380]$ in Figure 2 involved the vector parameter form (2.17) of the general accuracy formula, carried out by the bootstrap sampling method of Section 3.

With $I_c(\gamma)$ the indicator function of $\gamma \le c$, 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 $(d_{\rm lo}, d_{\rm up}) = (0.466, 0.330)$ at the interval endpoints;

$$\begin{pmatrix} d_{\rm lo} & 0 \\ 0 & d_{\rm up} \end{pmatrix}^{-1} \widehat{\rm var} \begin{pmatrix} d_{\rm lo} & 0 \\ 0 & d_{\rm up} \end{pmatrix} = \begin{pmatrix} .0476 & .0678 \\ .0678 & .0968 \end{pmatrix} \tag{6.26}$$

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

Remark 7. *Abc 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 $\hat{\theta} \pm c \cdot \widehat{\rm sd}$. An R version of *abc* is available from the author.

The solid bars in Figure 3 are abc intervals for the point estimates

$$\hat{\theta}_c = \widehat{\Pr}\left\{\gamma_{125} \le c | \hat{\beta}\right\} \tag{6.27}$$

(2.26). Here $\hat{G}$ (4.10) was the p-parameter exponential family, $p = 10$, with α_i (2.23) the $B = 10,000$ MCMC vectors, weights $p_i = 1$ in (4.8). Taking $\hat{G}$'s reversed roles of α and β into consideration, the *abc* call was

$$\texttt{abc(TT, ahat, S, bhat, mu)} \tag{6.28}$$

where `mu` was the function

$$\texttt{mu}(b) = \sum_{i=1}^{B} W_i(b)\alpha_i \Big/ \sum_{i=1}^{B} W_i(\beta) \qquad \left[W_i(b) = e^{(b-\hat{\beta})^\top \alpha_i}\right], \tag{6.29}$$

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

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

where $t_{ci} = t_c(\alpha_i)$ (2.28), while $\texttt{mu}^{-1}(\cdot)$ was the inverse function of $\texttt{mu}(\cdot)$, calculated to accuracy 10^{-11} 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 5 displays a portion of the *abc* output going into Figure 3. Besides $\hat{\theta}$ and $\widehat{\rm sd}$, it shows the three second-order correction coefficients 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, 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 Figure 3 involve one nonlocal recomputation. They enjoy tranformation invariance, monotone transformations of the parameter of interest producing the same transformation of the interval endpoints, which might be helpful for parameters like θ_c restricted to interval $[0, 1]$. However in this case they were not much different than the *abcq* versions.

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 Figure 6. *Tweedie's formula* (Efron, 2011) says that

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

Table 5: *abc* calculations for the diabetes data, Figure 3; (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, columns 7 and 8, were not much different than the purely local *abcq* limits, columns 9 and 10.

						abc		abcq	
c	$\hat{\theta}$	$\widehat{\text{sd}}$	a	z_0	c_q	lo	up	lo	up
.04	.00	.01	.00	.27	1.50	.00	.06	.00	.03
.08	.01	.03	.01	.21	1.17	.00	.13	.01	.09
.12	.04	.08	.00	.12	.88	.00	.25	.02	.22
.16	.11	.19	.00	.05	.60	.02	.44	.03	.45
.2	.25	.32	.00	.02	.33	.05	.63	.04	.68
.24	.46	.40	.00	−.02	.05	.13	.80	.08	.86
.28	.67	.36	.00	−.02	−.23	.28	.92	.23	.95
.32	.84	.24	.00	−.03	−.50	.49	.98	.47	.96
.36	.94	.12	.00	−.03	−.78	.71	.99	.73	.97

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

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

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

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

Theoretically there is a version of Tweedie's formula applying to any function $\tau(\delta)$ (called "Bayes rule in terms of f" in Efron, 2013). The case $\tau(\delta) = \delta$, however, is particularly favorable to GLM modeling of the marginal density $f(z)$ (5.4). Other choices of $\tau(\delta)$ may require non-GLM models for f, returning hierarchical Bayes analysis to the general, nonexponential family framework of Section 2.

Remark 9. *Empirical Bayes sd 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 $\boldsymbol{f}_\alpha$. Let $\boldsymbol{Q}$ be the gradient of $t(\alpha) = E\{\tau(\delta_0)|z_0, \alpha\}$ with respect to $\boldsymbol{f}$ (specific formulas for Q are given in Efron, 2013), and $\dot{\boldsymbol{f}}_\alpha$ the $J \times p$ derivative matrix $(\partial \boldsymbol{f}_{\alpha_j}/\partial\alpha_k)$. Then a first-order Taylor expansion gives

$$t(\hat{\alpha}) - t(\alpha) \doteq \boldsymbol{Q}^\top \dot{\boldsymbol{f}}_\alpha(\hat{\alpha} - \alpha). \tag{6.34}$$

This yields

$$\text{sd}\left(t(\hat{\alpha})\right)^2 \doteq \boldsymbol{Q}^\top \dot{\boldsymbol{f}}_\alpha \Sigma_\alpha \dot{\boldsymbol{f}}_\alpha \boldsymbol{Q} \qquad [\Sigma_\alpha = \text{cov}_\alpha(\hat{\alpha})] \tag{6.35}$$

and

$$\text{cov}\left(t(\hat{\alpha}), \hat{\alpha}\right) \doteq \boldsymbol{Q}^\top \dot{\boldsymbol{f}}_\alpha \Sigma_\alpha, \tag{6.36}$$

so

$$\text{sd}\left(t(\hat{\alpha})\right)^2 \doteq \text{cov}\left(t(\hat{\alpha}), \hat{\alpha}\right)^\top \Sigma_\alpha^{-1} \text{cov}\left(t(\hat{\alpha}), \hat{\alpha}\right). \tag{6.37}$$

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

$$\text{sd}\left(t(\hat{\alpha})\right)^2 \doteq \text{cov}\left(t(\hat{\alpha}), \hat{\alpha}\right)^\top V_\alpha \text{cov}\left(t(\hat{\alpha}), \hat{\alpha}\right). \tag{6.38}$$

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

References

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Anal.* 1: 385–402 (electronic).

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. New York: Springer-Verlag, 2nd ed.

Carlin, B. P. and Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. Texts in Statistical Science. Chapman & Hall/CRC, 2nd ed.

Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates. *Ann. Statist.* 14: 1–67, with a discussion and a rejoinder by the authors.

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. *J. Amer. Statist. Assoc.* 82: 171–200, with comments and a rejoinder by the author.

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

Efron, B. (2012). Bayesian inference and the parametric bootstrap. *Ann. Appl. Statist.* 6: 1971–1997.

Efron, B. (2013). Two modeling strategies for empirical Bayes estimation. *Statist. Sci.* Submitted.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* 32: 407–499, with discussion, and a rejoinder by the authors.

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. *J. Roy. Statist. Soc. Ser. B* 74: 419–474.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995). *Bayesian Data Analysis*. Texts in Statistical Science Series. London: Chapman & Hall.

Ghosh, M. (2011). Objective priors: An introduction for frequentists. *Statist. Sci.* 26: 187–202, with discussion and a rejoinder by the author.

Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *J. Amer. Statist. Assoc.* 91: 1343–1370.

Little, R. J. (2006). Calibrated Bayes: A Bayes/frequentist roadmap. *Amer. Statist.* 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. *Commun. Statist.-Simulat. Comput.* 19: 837–845.

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

Park, T. and Casella, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.* 103: 681–686.

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.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* 58: 267–288.

Welch, B. L. and Peers, H. W. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *J. Roy. Statist. Soc. Ser. B* 25: 318–329.