BAYESIAN INFERENCE AND THE PARAMETRIC BOOTSTRAP

By

Bradley Efron

Technical Report 258
November 2011

Division of Biostatistics
STANFORD UNIVERSITY
Stanford, California
Bayesian inference and the parametric bootstrap

Bradley Efron*
Stanford University

Abstract

The parametric bootstrap can be used for the efficient computation of Bayes posterior distributions. Importance sampling formulas take on an easy form relating to the deviance in exponential families, and are particularly simple starting from Jeffreys invariant prior. Because of the i.i.d. nature of bootstrap sampling, familiar formulas describe the computational accuracy of the Bayes estimates. Besides computational methods, the theory provides a connection between Bayesian and frequentist analysis. Efficient algorithms for the frequentist accuracy of Bayesian inferences are developed and demonstrated in a model selection example.

Keywords: Jeffreys prior, exponential families, deviance, generalized linear models

1 Introduction

This article concerns the use of the parametric bootstrap to carry out Bayesian inference calculations. Two main points are made: that in those cases where bootstrap methods apply, they offer an efficient and computationally straightforward way to compute posterior distributions and estimates, enjoying some advantages over Markov chain techniques; and that the parametric bootstrap helps connect Bayes and frequentist points of view.

The basic idea is simple and not unfamiliar: that the bootstrap is useful for importance sampling computation of Bayes posterior distributions. An important paper by Newton and Raftery (1994) suggested a version of non-parametric bootstrapping for this purpose. By “going parametric” we can make the Bayes/bootstrap connection more transparent.

Section 2 introduces the main ideas in terms of an elementary one-parameter example, and illustrates a connection between Jeffreys invariant prior density and second-order accurate bootstrap confidence limits. Both methods are carried out via reweighting of the original “raw” bootstrap replications. The calculation of posterior distributions by bootstrap reweighting is a main theme here, in constrast to Markov chain methods, which strive to directly produce correctly distributed posterior realizations.

Multidimensional exponential families, discussed in Section 3, allow the Bayes/bootstrap conversion process to be explicitly characterized. Two important families, multivariate normal and generalized linear models, are investigated in Section 4 and Section 5. Jeffreys-type priors can yield unsatisfactory results in multiparameter problems (Ghosh, 2011), as shown here by comparison with bootstrap confidence limits.

An advantage of bootstrap reweighting schemes is the straightforward analysis of their accuracy. Section 6 develops accuracy estimates for our methodology, both internal (How

*Research supported in part by NIH grant 8R01 EB002784 and by NSF grant DMS 0804324.
many bootstrap replications are necessary?) and external (How much would the results vary
in future data sets?). The latter concerns the frequentist analysis of Bayesian estimates, an
important question in “objective Bayes” applications; see for instance Gelman, Meng and Stern

Bootstrap reweighting can apply to any choice of prior (not favoring convenience priors such
as the conjugates, for example), but here we will be most interested in the objective-type Bayes
analyses that dominate current practice. Jeffreys priors are featured in the examples, more for
easy presentation than necessity. The paper ends with a brief summary in Section 7. Some
technical details are deferred to the Appendix.

Connections between nonparametric bootstrapping and Bayesian inference emerged early,
with the “Bayesian bootstrap,” Rubin (1981) and Efron (1982). Bootstrap reweighting is
deployed differently in Smith and Gelfand (1992), with a nice example given in their Section 5.
Sections 4 and 6 of Efron and Tibshirani (1998) develop bootstrap reweighting along the lines
used in this paper.

2 Conversion and reweighting

Our methodology is introduced here in terms of a simple one-parameter problem. Table 1 shows
scores for \( n = 22 \) students on two tests, “mechanics” and “vectors”, having sample correlation
\[
\hat{\theta} = 0.498.
\]  

(2.1)

We wish to calculate some measure of posterior distribution for the true underlying parameter
value
\[
\theta_0 = \text{correlation (mechanics score, vectors score).}
\]  

(2.2)

Table 1: Scores of 22 students on two tests, “mechanics” and “vectors” (from Mardia et al., 1979, a
randomly chosen subset of the 88 students in their Table 1.2.1). The sample correlation is \( \hat{\theta} = 0.498 \).

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
</tr>
</thead>
<tbody>
<tr>
<td>mech</td>
<td>7</td>
<td>44</td>
<td>49</td>
<td>59</td>
<td>34</td>
<td>46</td>
<td>0</td>
<td>32</td>
<td>49</td>
<td>52</td>
<td>44</td>
<td>36</td>
<td>42</td>
<td>5</td>
<td>22</td>
<td>18</td>
<td>41</td>
<td>48</td>
<td>31</td>
<td>42</td>
<td>46</td>
<td>63</td>
</tr>
<tr>
<td>vec</td>
<td>51</td>
<td>69</td>
<td>41</td>
<td>70</td>
<td>42</td>
<td>40</td>
<td>40</td>
<td>45</td>
<td>57</td>
<td>64</td>
<td>59</td>
<td>60</td>
<td>58</td>
<td>30</td>
<td>58</td>
<td>51</td>
<td>63</td>
<td>38</td>
<td>42</td>
<td>69</td>
<td>49</td>
<td>63</td>
</tr>
</tbody>
</table>

As in Mardia et al. (1979), we assume that the individual student scores \( y_i = (\text{mech}_i, \text{vec}_i) \)
are a random sample from a bivariate normal distribution having unknown mean vector \( \mu \) and
covariance matrix \( \Sigma \),
\[
y : y_i \overset{\text{ind}}{\sim} N_2(\mu, \Sigma) \quad \text{for } i = 1, 2, \ldots, 22,
\]  

(2.3)

with \( y = (y_1, y_2, \ldots, y_{22}) \) representing the full data set. Let \((\hat{\mu}, \hat{\Sigma})\) denote the usual maximum
likelihood estimate (MLE). Then a \textit{parametric bootstrap sample} \( y^* \) follows (2.3), with \((\hat{\mu}, \hat{\Sigma})\)
replacing \((\mu, \Sigma)\),
\[
y^* : y_i^* \overset{\text{ind}}{\sim} N_2(\hat{\mu}, \hat{\Sigma}) \quad \text{for } i = 1, 2, \ldots, 22.
\]  

(2.4)

The sample correlation of \( y^* \) is a \textit{parametric bootstrap replication} of \( \hat{\theta} \), say \( \hat{\theta}^* \). A total of
\( B = 10,000 \) parametric bootstrap samples \( y^* \) were independently generated according to (2.4),
and the corresponding \( \hat{\theta}^* \) values calculated. We will denote them simply as
\[
\theta_1, \theta_2, \ldots, \theta_i, \ldots, \theta_B,
\] (2.5)
with \( \theta_i \) short for \( \hat{\theta}_i^* \).

Figure 1: Histogram of \( B = 10,000 \) bootstrap replications for the student score correlation coefficient (2.4)–(2.5) scaled to integrate to 1. Solid curve is Fisher’s density formula (2.6) for \( \theta = 0.498 \). Triangles indicate the exact 95% confidence interval \( \theta \in (0.093, 0.741) \).

The histogram in Figure 1 compares the distribution of the 10,000 \( \theta_i \)’s with Fisher’s theoretical density function \( f_\theta(\hat{\theta}) \),
\[
f_\theta(\hat{\theta}) = \frac{(n-2)(1-\theta^2)^{(n-1)/2}(1-\hat{\theta}^2)^{(n-4)/2}}{\pi} \int_0^\infty dw \frac{w}{(\cosh w - \theta \hat{\theta})^{n-1}} \] (2.6)
where \( \theta \) has been set equal to its MLE value 0.498. In this sense \( f_{0.498}(\cdot) \) is the ideal parametric bootstrap density we would obtain if the number of replications \( B \) approached infinity. Chapter 32 of Johnson and Kotz (1970) gives formula (2.6) and other representations of \( f_\theta(\hat{\theta}) \).

Figure 1 also indicates the exact 95% confidence limits
\[
\theta_0 \in (0.093, 0.741),
\] (2.7)
2\( 1/2 \)\% noncoverage in each tail, obtained from \( f_\theta(\hat{\theta}) \) by the usual construction,
\[
\int_{0.093}^1 f_{0.498}(\theta) \ d\theta = 0.025,
\] (2.8)
and similarly at the upper endpoint.

Suppose now we have a prior density \( \pi(\theta) \) for the parameter \( \theta \), and wish to calculate the posterior density \( \pi(\theta|\hat{\theta}) \). For any subset \( \mathcal{A} \) of the parameter space \( \Theta = [-1, 1] \),
\[
\Pr \left\{ \theta \in \mathcal{A} | \hat{\theta} \right\} = \frac{\int_{\mathcal{A}} \pi(\theta) f_\theta(\hat{\theta}) \ d\theta}{\int_\Theta \pi(\theta) f_\theta(\hat{\theta}) \ d\theta}
\] (2.9)
according to Bayes rule.

Define the conversion factor $R(\theta)$ to be the ratio of the likelihood function to the bootstrap density,

$$R(\theta) = \frac{f_\theta(\hat{\theta})}{f_\hat{\theta}(\theta)}.$$  \hfill (2.10)

Here $\hat{\theta}$ is fixed at its observed value 0.498 while $\theta$ represents any point in $\Theta$. We can rewrite (2.9) as

$$\Pr\{\theta \in A|\hat{\theta}\} = \frac{\int_A \pi(\theta) R(\theta) f_\hat{\theta}(\theta) d\theta}{\int_{\Theta} \pi(\theta) R(\theta) f_\hat{\theta}(\theta) d\theta}. \hfill (2.11)$$

More generally, if $t(\theta)$ is any function of $\theta$ its posterior expectation is

$$E\{t(\theta)|\hat{\theta}\} = \frac{\int_{\Theta} t(\theta) \pi(\theta) R(\theta) f_\hat{\theta}(\theta) d\theta}{\int_{\Theta} \pi(\theta) R(\theta) f_\hat{\theta}(\theta) d\theta}. \hfill (2.12)$$

The integrals in (2.11) and (2.12) are now being taken with respect to the parametric bootstrap density $f_\hat{\theta}(\cdot)$. Since $\theta_1, \theta_2, \ldots, \theta_B$ (2.5) is a random sample from $f_\hat{\theta}(\cdot)$, the integrals can be estimated by sample averages in the usual way, yielding the familiar importance sampling estimate of $E\{t(\theta)|\hat{\theta}\}$,

$$\hat{E}\{t(\theta)|\hat{\theta}\} = \frac{\sum_{i=1}^{B} t_i \pi_i R_i}{\sum_{i=1}^{B} \pi_i R_i} \hfill (2.13)$$

where $t_i = t(\theta_i)$, $\pi_i = \pi(\theta_i)$, and $R_i = R(\theta_i)$. Under mild regularity conditions, the law of large numbers implies that $\hat{E}\{t(\theta)|\hat{\theta}\}$ approaches $E\{t(\theta)|\theta\}$ as $B \to \infty$. (The accuracy calculations of Section 6 will show that in this case $B = 10,000$ was vastly larger than necessary for most purposes.)

![Figure 2: Heavy curve is the posterior density $\pi(\theta|\hat{\theta})$ for the correlation (2.2), starting from Jeffreys prior (2.14), obtained by reweighting the $B = 10,000$ bootstrap replications (2.5); triangles show 95% credible limits $\theta_0 \in (0.095, 0.748)$. Light dashed curve is raw unweighted bootstrap distribution. Beaded curve is BCa weighted bootstrap density (2.17), nearly the same as $\pi(\hat{\theta}|\theta)$ in this case.](image)

The heavy curve in Figure 2 describes $\hat{\pi}(\theta|\hat{\theta})$, the estimated posterior density starting from Jeffreys prior

$$\pi(\theta) = \frac{1}{1 - \theta^2}. \hfill (2.14)$$
(see Section 3). The raw bootstrap distribution puts weight $1/B$ on each of the $B$ replications $\theta_i$. By reweighting these points proportionately to $w_i = \pi_i R_i$ we obtain the estimated posterior distribution of $\theta$ given $\hat{\theta}$, with

$$\Pr\{\theta \in A|\hat{\theta}\} = \frac{\sum_{i=1}^{B} w_i}{w_i \sum_{i=1}^{B} w_i; \quad (2.15)}$$

$\hat{\pi}(\theta|\hat{\theta})$ represents the density of this distribution — essentially a smoothed histogram of the 10,000 $\theta_i$’s, weighted proportionally to $w_i$.

Integrating $\hat{\pi}(\theta|\hat{\theta})$ yields the 95% credible limits (2/2% posterior probability in each tail)

$$\theta_0 \in (0.095, 0.748), \quad (2.16)$$
close to the exact limits (2.7). Prior (2.14) is known to yield accurate frequentist coverage probabilities, being a member of the Welch–Peers family discussed in Section 4.

In this case, the weights $w_i = \pi_i R_i$ can be thought of as correcting the raw unweighted ($w_i \equiv 1$) bootstrap density. Figure 2 shows the correction as a small shift leftwards. BCa, standing for bias-corrected and accelerated, is another set of corrective weights, obtained from purely frequentist considerations. Letting $\hat{G}(\theta)$ denote the usual empirical cumulative distribution function (cdf) of the bootstrap replications $\theta_1, \theta_2, ..., \theta_B$, the BCa weight on $\theta_i$ is

$$w_{BCa} = \frac{\varphi\left(z_{\hat{\theta}i}/(1 + az_{\hat{\theta}i}) - z_0\right)}{(1 + az_{\hat{\theta}i})^2 \varphi(z_{\hat{\theta}i} + z_0)} \left[z_{\hat{\theta}i} = \Phi^{-1}\hat{G}(\theta_i) - z_0\right] \quad (2.17)$$

where $\varphi$ and $\Phi$ are the standard normal density and cdf, while $z_0$ and $a$ are the bias-correction and acceleration constants developed in Efron (1987) and DiCiccio and Efron (1992), further discussed in Section 4 and the Appendix. Their estimated values are $z_0 = -0.068$ and $a = 0$ for the student score correlation problem. The BCa density $\pi_{BCa}(\hat{\theta}|\theta)$, obtained by reweighting as in (2.15), is seen in Figure 2 to nicely agree with the Jeffreys posterior density, being slightly heavier in the left tail, with 95% central interval $\theta_0 \in (0.074, 0.748)$. This agreement is misleadingly soothing, as will be seen in the multidimensional context of Section 4.

### 3 Exponential families

The Bayes/bootstrap conversion process takes on a simplified form in exponential families. This facilitates its application to multiparameter problems, as discussed here and in the next two sections.

The density functions for a $p$-parameter exponential family $\mathcal{F}$ can be expressed as

$$f_\beta(\hat{\beta}) = e^{\alpha’ \hat{\beta} - \psi(\alpha)} f_0(\hat{\beta}) \quad (3.1)$$

where the $p$-vector $\alpha$ is the canonical parameter, $\hat{\beta}$ is the $p$-dimensional sufficient statistic vector, and where $\psi(\alpha)$, the cumulant generating function, provides the multipliers necessary for $f_\beta(\hat{\beta})$ integrating to 1. Here we have indexed the family by its expectation parameter vector $\beta$,

$$\beta = E_\alpha(\hat{\beta}) \quad (3.2)$$
for the sake of subsequent notation, but \( \alpha \) and \( \beta \) are one-to-one functions, and we could just as well write \( f_\alpha(\hat{\beta}) \).

The \textit{deviance} between any two members of \( \mathcal{F} \) is

\[
D(\beta_1, \beta_2) = 2E_{\beta_1} \left\{ \log \left( \frac{f_{\beta_1}(\hat{\beta})}{f_{\beta_2}(\hat{\beta})} \right) \right\}
\]  

(3.3)

(denoted equivalently \( D(\alpha_1, \alpha_2) \) since deviance does not depend on the parameterization of \( \mathcal{F} \).) Taking logs in (3.1) shows that

\[
D(\beta_1, \beta_2) = (\alpha_1 - \alpha_2)' \beta_1 - (\psi(\alpha_1) - \psi(\alpha_2)).
\]  

(3.4)

Then family (3.1) can be re-expressed in “Hoeffding’s form” as

\[
f_{\beta}(\hat{\beta}) = f_{\hat{\beta}}(\hat{\beta}) e^{-D(\hat{\beta}, \beta)/2}.
\]  

(3.5)

Since \( D(\hat{\beta}, \beta) \) is equal or greater than zero, (3.5) shows that \( \beta = \hat{\beta} \) is the MLE, maximizing \( f_{\beta}(\hat{\beta}) \) over all choices of \( \beta \) in \( \mathcal{B} \), the space of possible expectation vectors.

\textit{Parametric bootstrap replications} of \( \hat{\beta} \) are independent draws from \( f_{\hat{\beta}}(\cdot) \),

\[
f_{\hat{\beta}}(\cdot) \longrightarrow \beta_1, \beta_2, \ldots, \beta_i, \ldots, \beta_B
\]  

(3.6)

where \( \beta_i \) is shorthand notation for \( \hat{\beta}_i^* \). Starting from a prior density \( \pi(\beta) \) on \( \mathcal{B} \), the posterior expectation of any function \( t(\beta) \) given \( \hat{\beta} \) is estimated by

\[
\hat{E} \left\{ t(\beta) | \hat{\beta} \right\} = \sum_{i=1}^{B} \frac{t(\beta_i) \pi(\beta_i) R(\beta_i)}{\sum_{i=1}^{B} \pi(\beta_i) R(\beta_i)}
\]  

(3.7)

as in (2.13), with \( R(\beta) \) the \textit{conversion factor}

\[
R(\beta) = f_{\hat{\beta}}(\hat{\beta}) / f_{\hat{\beta}}(\beta)
\]  

(3.8)

\textit{Note:} \( \pi(\beta) R(\beta) \) is transformation invariant, so formula (3.7) produces the same numerical result if we bootstrap \( \alpha_1, \alpha_2, \ldots, \alpha_B \) instead of (3.6), or for that matter bootstrap any other sufficient vector. See Section 4.

Hoeffding’s form (3.5) allows a convenient expression for \( R(\beta) \):

\textbf{Lemma 1.} \textit{Conversion factor} (3.8) \textit{equals}

\[
R(\beta) = \xi(\beta) e^{\Delta(\beta)}
\]  

(3.9)

where

\[
\xi(\beta) = f_{\hat{\beta}}(\hat{\beta}) / f_{\hat{\beta}}(\beta)
\]  

(3.10)

and

\[
\Delta(\beta) = \left[ D(\beta, \hat{\beta}) - D(\hat{\beta}, \beta) \right] / 2.
\]  

(3.11)
Letting $\hat{\alpha}$ be the canonical parameter vector corresponding to $\hat{\beta}$, (3.4) gives
\[
\Delta(\beta) = (\alpha - \hat{\alpha})' \left( \beta + \hat{\beta} \right) - 2 [\psi(\alpha) - \psi(\hat{\alpha})],
\]
which is useful for both theoretical and numerical computations.

The derivatives of $\psi$ with respect to components of $\alpha$ yield the moments of $\hat{\beta}$,
\[
\hat{\psi}(\alpha) \equiv (\partial \psi / \partial \alpha_j) = \beta, \quad \hat{\psi}(\alpha) \equiv (\partial^2 \psi / \partial \alpha_j \partial \alpha_k) = V(\alpha) = \text{cov}_\alpha \left\{ \hat{\beta} \right\},
\]
and
\[
\hat{\psi}(\alpha) \equiv (\partial^3 \psi / \partial \alpha_j \partial \alpha_k \partial \alpha_l) = U(\alpha),
\]

In repeated sampling situations, where $\hat{\beta}$ is obtained from $n$ independent observations, the entries of $V(\alpha)$ and $U(\alpha)$ are typically of order $O(n^{-1})$ and $O(n^{-2})$, respectively; see Section 5 of Efron (1987).

The normal approximation
\[
\hat{\beta} \sim \mathcal{N}_p(\beta, V(\alpha))
\]
yields
\[
f_{\hat{\beta}}(\beta) \doteq (2\pi)^{-n/2}|V(\alpha)|^{-1/2} \quad \text{and} \quad f_{\hat{\beta}} \left( \hat{\beta} \right) \doteq (2\pi)^{-n/2}|V(\hat{\alpha})|^{-1/2},
\]
so
\[
\xi(\beta) \doteq |V(\alpha)|^{1/2} / |V(\hat{\alpha})|^{1/2}.
\]
Because (3.16) applies the central limit theorem where it is most accurate, at the center, (3.17) typically errs by a factor of only $1 + O(1/n)$ in repeated sampling situations; see Tierney and Kadane (1986). In fact for discrete families like the Poisson, where $\hat{\beta}$ is obtained from independent observations, the entries of $V(\alpha)$ and $U(\alpha)$ are typically of order $O(n^{-1})$ and $O(n^{-2})$, respectively; see Section 5 of Efron (1987).

Jeffreys invariant prior density, as described in Kass and Wasserman (1996), takes the form
\[
\pi^{\text{Jeff}}(\beta) = c |V(\alpha)|^{-1/2}
\]
in family (3.1), with $c$ an arbitrary positive constant that does not affect estimates such as (3.7). Ignoring $c$, we can use (3.17)–(3.18) to rewrite the conversion factor $R(\beta)$ (3.9) as
\[
R(\beta) = e^{\Delta(\beta) / \pi^{\text{Jeff}}(\beta)}.
\]

At this point it is convenient to define the adjusted prior and adjusted conversion factor
\[
\pi^+(\beta) = \pi(\beta) / \pi^{\text{Jeff}}(\beta) \quad \text{and} \quad R^+(\beta) = R(\beta) \cdot \pi^{\text{Jeff}}(\beta) = e^{\Delta(\beta)};
\]
\[
\pi^+(\beta) R^+(\beta) \text{ equals } \pi(\beta) R(\beta), \text{ and can be used just as well for estimating } \hat{E}\{t(\beta) | \hat{\beta} \} \text{ in (3.7).}
\]

Jeffreys prior is intended to be “uninformative.” Like other objective priors discussed in Kass and Wasserman, it is designed for Bayesian use in situations lacking prior experience. Its use amounts to choosing $\pi^+(\beta)$ to be flat, $\pi^+(\beta) = 1$, in which case (3.7) takes on a particularly simple form:

**Lemma 2.** If $\pi(\beta)$ is Jeffreys prior (3.18), then (3.7) equals
\[
\hat{E}\{t(\beta) | \hat{\beta} \} = \sum_{i=1}^{B} t(\beta_i) e^{\Delta(\beta_i)} / \sum_{i=1}^{B} e^{\Delta(\beta_i)}
\]
with $\Delta(\beta)$ as in (3.11)–(3.12).
The normal translation model $\hat{\beta} \sim \mathcal{N}_p(\beta, \Sigma)$, with $\Sigma$ fixed, has $\Delta(\beta) = 0$, so that the Bayes estimate $\hat{t}$ in (3.21) equals the unweighted bootstrap estimate $\bar{t}$,

$$\hat{t} = \mathbb{E}\left\{ t(\beta) | \hat{\beta} \right\} = \frac{\sum_{i=1}^{B} t_i}{B} = \bar{t}. \quad (3.22)$$

Usually though $\hat{t}$ will not equal $\bar{t}$, the difference relating to the variability of $\Delta(\beta)$ in (3.21).

A simple but informative result concerns the relative Bayesian difference (RBD) of $t(\beta)$ defined to be

$$\text{RBD}(t) = \frac{(\hat{t} - \bar{t})}{\text{sd}(t)}, \quad (3.23)$$

where $\text{sd}(t) = \left[ \frac{\sum_{i=1}^{B} (t_i - \bar{t})^2}{B} \right]^{1/2}$.

**Lemma 3.** Letting $r_i = \pi_i R_i$, the relative Bayesian difference of $t(\beta)$ is

$$\text{RBD}(t) = \text{cor}(t, r) \cdot \text{cv}(r), \quad (3.24)$$

and if $\pi(\beta) = \pi^{\text{Jeff}}(\beta)$,

$$\text{RBD}(t) = \text{cor}(t, r) \cdot \text{sd}(\Delta); \quad (3.25)$$

here $\text{cor}(t, r)$ is the empirical correlation between $t_i$ and $r_i$ for the $B$ bootstrap replications, $\text{cv}(r)$ the empirical coefficient of variation of the $r_i$ values, and $\text{sd}(\Delta)$ the empirical standard deviation of the $\Delta_i$ values.

**Proof.** (3.24) follows immediately from (3.7),

$$\text{RBD}(t) = \frac{\sum_{i=1}^{B} (t_i - \hat{t}) r_i / B}{\text{sd}(t) \sum_{i=1}^{B} r_i / B} = \text{cor}(t, r) \frac{\text{sd}(r)}{\bar{r}}. \quad (3.26)$$

If $\pi(\beta)$ is the Jeffreys prior (3.18) then $r(\beta) = \exp(\Delta(\beta))$ (3.19), and the usual delta-method argument gives $\text{cv}(r) \approx \text{sd}(\Delta)$.

The student score example of Figure 2 (which isn’t in exponential family form) has, directly from definition (3.23),

$$\text{RBD}(t) = \frac{0.473 - 0.490}{0.169} = -0.101, \quad (3.27)$$

which is also obtained from (3.24) with $\text{cor}(t, r) = -0.945$ and $\text{cv}(r) = 0.108$. Notice that the $\text{cor}(t, r)$ factor in (3.24), and likewise $\text{sd}(\Delta)$ in (3.25), apply to any function $t(\beta)$, only the $\text{cor}(t, r)$ factor being particular. The multiparameter examples of Sections 3 and 4 have larger $\text{cv}(r)$ but smaller $\text{cor}(t, r)$, again yielding rather small values of RBD($t$). All of the Jeffreys prior examples in this paper show substantial agreement between the Bayes and unweighted bootstrap results.

Asymptotically, the deviance difference $\Delta(\beta)$ depends on the skewness of the exponential family. A normal translation family has zero skewness, with $\Delta(\beta) = 0$ and $R(\beta) = 1$, so the unweighted parametric bootstrap distribution is the same as the flat-prior Bayes posterior distribution. In a repeated sampling situation, skewness goes to zero as $n^{-1/2}$, making the Bayes and bootstrap distributions converge at this rate. We can provide a simple statement in one-parameter families:
Theorem 1. In a one-parameter exponential family, $\Delta(\beta)$ has the Taylor series approximation

$$\Delta(\beta) \approx \frac{1}{6} \hat{\gamma} Z^3 \quad \left[ Z = \hat{V}^{-1/2} \left( \beta - \hat{\beta} \right) \right]$$

(3.28)

where $\hat{V}$ and $\hat{\gamma}$ are the variance and skewness of $\beta \sim f_{\hat{\beta}}(\cdot)$. In large-sample situations, $Z \sim N(0, 1)$ and $\hat{\gamma}$ is $O(n^{-1/2})$, making $\Delta(\beta)$ of order $O_p(n^{-1/2})$.

(Proof appears in the Appendix, along with the theorem’s multiparameter version.)

As a simple example, suppose

$$\hat{\beta} \sim \beta \text{ Gamma}_n/n \quad [\beta \in (0, \infty)],$$

(3.29)

so $\hat{\beta}$ is a scaled version of a standard Gamma variate having $n$ degrees of freedom. In this case,

$$\Delta(\beta) \approx \frac{1}{3 \sqrt{n}} Z^3 \quad \text{with} \quad Z = \sqrt{n} \left( \frac{\beta}{\hat{\beta}} - 1 \right),$$

(3.30)

making $\Delta(\beta)$ an increasing cubic function of $\beta$. The cubic nature of (3.28) and (3.30) makes reweighting of the parametric bootstrap replications $\beta_i$ by $\exp(\Delta_i)$ more extreme in the tails of the distribution than near $\hat{\beta}$.

Stating things in terms of conditional expectations $\hat{E}\{t(\beta)|\hat{\beta}\}$ as in (3.7) is convenient, but partially obscures the basic idea: that the distribution putting weight proportional to $w_i = \pi_i R_i$ on $\beta_i$ approximates the posterior distribution $\pi(\beta|\hat{\beta})$.

As an example of more general Bayesian calculations, consider the “posterior predictive distribution,”

$$g(y) = \int \pi(\beta|\hat{\beta}) g_\beta(y) \, d\beta$$

(3.31)

where $y$ is the original data set yielding $\hat{\beta}$ by sufficiency, as in (2.3), having density functions $g_\beta(y) = f_\beta(\hat{\beta}) h(y|\hat{\beta})$. For each $\beta_i$ we sample $y_i^{*\ast}$ from $g_\beta(\cdot)$. Then the discrete distribution putting weight proportional to $w_i$ on $y_i^{*\ast}$, for $i = 1, 2, \ldots, B$, approximates $g(y)$. See Gelman et al. (1996).

4 The multivariate normal family

This section and the next illustrate Bayes/bootstrap relationships in two important exponential families: the multivariate normal and generalized linear models. A multivariate normal sample $y$ comprises $n$ independent $d$-dimensional normal vector observations

$$y : y_i \overset{\text{ind}}{\sim} N_d(\mu, \Sigma) \quad i = 1, 2, \ldots, n.$$  

(4.1)

This involves $p = d \cdot (d + 3)/2$ unknown parameters, $d$ for the mean vector $\mu$ and $d \cdot (d + 1)/2$ for the covariance matrix $\Sigma$. We will use $\gamma$ to denote the vector of all $p$ parameters; $\gamma$ is not the expectation vector $\beta$ (3.2), but rather a one-to-one quadratic function $\gamma = m(\beta)$ described in formula (3.5) of DiCiccio and Efron (1992).

The results of Section 3 continue to hold under smooth one-to-one transformations $\gamma = m(\beta)$. Let $\hat{f}_\gamma(\hat{\gamma})$ denote the density of the MLE $\hat{\gamma} = m(\hat{\beta})$, and likewise $R(\gamma) = \hat{f}_\gamma(\hat{\gamma})/\hat{f}_\gamma(\gamma)$ for the conversion factor, $D(\gamma_1, \gamma_2)$ for the deviance, $\hat{\Delta}(\gamma) = [D(\gamma, \hat{\gamma}) - D(\hat{\gamma}, \gamma)]/2$ for the deviance
difference, and $\tilde{\pi}_{\text{Jeff}}(\gamma)$ for Jeffreys prior. Then Lemma 1 continues to apply in the transformed coordinates:

$$
\tilde{R}(\gamma) = \tilde{\xi}(\gamma) e^{\Delta(\gamma)}
$$

(4.2)

(See the Appendix.)

A parametric bootstrap sample

$$
\tilde{f}(\cdot) \longrightarrow \gamma_1, \gamma_2, \ldots, \gamma_B
$$

(4.3)

approximates the conditional expectation of a function $\tilde{t}(\gamma)$, starting from prior $\tilde{\pi}(\gamma)$, by

$$
\hat{E}\{\tilde{t}(\gamma)|\hat{\gamma}\} = \frac{\sum_{i=1}^{B} \tilde{t}_i \tilde{\pi}_i \tilde{R}_i}{\sum_{i=1}^{B} \tilde{\pi}_i \tilde{R}_i}
$$

as in (2.14), and if $\tilde{\pi}(\gamma)$ is Jeffreys prior,

$$
\hat{E}\{\tilde{t}(\gamma)|\hat{\gamma}\} = \frac{\sum_{i=1}^{B} \tilde{t}_i e^{\Delta_i}}{\sum_{i=1}^{B} e^{\Delta_i}}
$$

as in (3.21). This can be particularly handy since $\Delta$ is transformation invariant and can be evaluated in any convenient set of coordinates, while $\tilde{\pi}_{\text{Jeff}}(\gamma)$ need not be calculated at all.

The following theorem provides $\tilde{\xi}(\gamma)$ and $\tilde{R}(\gamma)$ for a multivariate normal sample (4.1), working with $\gamma$ the $p = d \cdot (d + 3)/2$ coordinates consisting of $\mu$ and the elements of $\Sigma$ on or above its main diagonal:

**Theorem 2.** In $(\mu, \Sigma)$ coordinates,

$$
\tilde{\xi}(\mu, \Sigma) = \left(\frac{|\Sigma|}{|\hat{\Sigma}|}\right)^{\frac{d+2}{2}}
$$

(4.6)

and

$$
\tilde{\Delta}(\mu, \Sigma) = n \left\{ (\mu - \hat{\mu})'\hat{\Sigma}^{-1} - \Sigma^{-1} (\mu - \hat{\mu}) + \frac{\text{tr}(\Sigma\Sigma^{-1} - \hat{\Sigma}\hat{\Sigma}^{-1})}{2} + \log \left| \frac{\hat{\Sigma}}{|\Sigma|} \right| \right\}.
$$

(4.7)

(Proof in the Appendix.)

Here $1/\tilde{\xi}(\mu, \Sigma)$ turns out to be exactly proportional to $|\hat{V}(\gamma)|^{-1/2}$, and either expression gives $\tilde{\pi}_{\text{Jeff}}(\mu, \Sigma)$. Expression (4.7) equals the deviance difference (3.11), no matter what the choice of coordinates.

Theorem 2 makes it easy to carry out parametric bootstrapping: having calculated the usual MLE estimates $(\hat{\mu}, \hat{\Sigma})$, each bootstrap data set $y^*$ is generated as in (4.1),

$$
y^*: y_i^* \sim N_d \left( \hat{\mu}, \hat{\Sigma} \right),
$$

(4.8)

from which we calculate the bootstrap MLE estimate $(\hat{\mu}^*, \hat{\Sigma}^*)$, denoted simply $(\mu, \Sigma)$ as before. To each of $B$ such replicates

$$(\mu, \Sigma)_1, (\mu, \Sigma)_2, \ldots, (\mu, \Sigma)_i, \ldots, (\mu, \Sigma)_B
$$

(4.9)
Figure 3: Heavy curve is Bayes posterior density for the eigenratio (4.11), starting from Jeffreys prior for a bivariate normal model; solid triangles show 95% credible limits (0.650, 0.908). Beaded curve is BCa confidence density based on weights (2.17) with $z_0 = -0.222$, $a = 0$; BCa 95% interval (0.598, 0.890), open triangles, is shifted far leftward. Light dashed curve is unweighted bootstrap density.

is attached the weight

$$w_i = \tilde{\pi}_i \xi_i e^{\tilde{\Delta}_i},$$

using Theorem 2 (or more exactly $w_i / \sum_{j=1}^B w_j$); this distribution, supported on the $B$ points (4.9), estimates the posterior distribution of $(\mu, \Sigma)$ given $(\hat{\mu}, \hat{\Sigma})$. Expectations are then obtained as in (4.4), and similarly for more general posterior parameters such as percentiles and credible limits.

Figure 3 applies this methodology to the student score data of Table 1, assuming the bivariate normal model (2.3). We take the parameter of interest $\theta$ to be the eigenratio

$$\theta = t(\mu, \Sigma) = \lambda_1 / (\lambda_1 + \lambda_2)$$

where $\lambda_1$ and $\lambda_2$ are the ordered eigenvalues of $\Sigma$; $\theta$ has MLE $\hat{\theta} = t(\hat{\mu}, \hat{\Sigma}) = 0.793$.

$B = 10,000$ bootstrap replications were generated as in (4.9), and $t_i = t((\mu, \Sigma)_i)$ calculated for each. Total computation time was about 30 seconds. The heavy curve shows the estimated posterior density of $\theta$ given $(\hat{\mu}, \hat{\Sigma})$, starting from Jeffreys prior. The 95% credible region, $2\frac{1}{2}$% probability excluded in each tail, was

Bayes: $\theta \in (0.650, 0.908)$.  

That is,

$$\sum_{t_i \leq 0.650} e^{\hat{\Delta}_i} / \sum_{j=1}^B e^{\hat{\Delta}_j} = 0.025,$$

and similarly for the upper endpoint.

In this case the BCa 95% confidence limits are shifted sharply leftward compared to (4.12),

BCa: $\theta \in (0.598, 0.890)$.  

11
The beaded curve in Figure 3 shows the full BCa confidence density, i.e., the estimated density based on the BCa weights (2.17). For the eigenratio, $z_0 = -0.222$ and $a = 0$ are the bias correction and acceleration constants. See the Appendix for a brief discussion of the $z_0$ and $a$ calculations.

Figure 4: Solid curve: BCa weights (2.17), with $(z_0, a) = (-0.222, 0)$, plotted versus bootstrap eigenratio replications $\theta_i$. Dashed curve: regression of Jeffreys prior Bayes weights $\exp(\tilde{\Delta}_i)$ on $\theta_\lambda$.

Figure 4 helps explain the difference between the Bayes and BCa results. The heavy curve shows the BCa weights (2.17) declining sharply as a function of $\theta_i = t((\mu, \Sigma)_i)$, the bootstrap eigenratio values. In other words, smaller values of $\theta_i$ are weighted more heavily, pulling the weighted percentile points of the BCa distribution downwards. On the other hand, the Bayes weights $\tilde{\pi}_i^{\text{Jeff}} \tilde{R}_i = \exp(\tilde{\Delta}_i)$ (represented in Figure 4 by their regression on $\theta_i$) are nearly flat, so that the Bayes posterior density is almost the same as the unweighted bootstrap density shown in Figure 3.

The BCa limits are known to yield highly accurate coverage probabilities; see DiCiccio and Efron (1996). Moreover, in the eigenratio case, the MLE $\hat{\theta}$ is strongly biased upwards, suggesting a downward shift for the confidence limits. This brings up a familiar complaint against Jeffreys priors, extensively discussed in Ghosh (2011): that in multiparameter settings they can give inaccurate inferences for individual parameters of interest.

This is likely to be the case for any general-purpose recipe for choosing objective prior distributions in several dimensions. For instance, repeating the eigenratio analysis with a standard inverse Wishart prior on $\Sigma$ (covariance matrix $I$, degrees of freedom 2) and a flat prior on $\mu$ gave essentially the same results as in Figure 3. Specific parameters of interest require specifically tailored priors, as with the Bernardo–Berger reference priors, again nicely reviewed by Ghosh (2011).

In fact, the BCa weights can be thought of as providing such tailoring: define the BCa prior (relative to the unweighted bootstrap distribution) to be

$$\pi_i^{\text{BCa}} = w_i^{\text{BCa}} / R_i$$

with $w_i^{\text{BCa}}$ as in (2.17). This makes the posterior weights $\pi_i^{\text{BCa}} R_i$ appearing in expressions like (3.7) equal the BCa weights $w_i^{\text{BCa}}$, and makes posterior credible limits based on the $\pi^{\text{BCa}}$ prior.
equal BCa limits. Formula (4.15) can be thought of as an automatic device for constructing Welch and Peers’ (1963) “probability matching priors”; see Tibshirani (1989).

5 Generalized linear models

The Bayes/bootstrap conversion theory of Section 3 applies directly to generalized linear models (GLM). A GLM begins with a one-parameter exponential family

\[ g_{\eta}(y) = e^{\eta y - \phi(\eta)} g_0(y) \]  \hspace{1cm} (5.1)

where \( \eta = \alpha \), \( y = \hat{\beta} \), and \( \phi(\eta) = \psi(\alpha) \) in notation (3.1). An \( n \times p \) structure matrix \( X \) and a \( p \)-dimensional parameter vector \( \alpha \) then yield an \( n \)-vector \( \eta = X\alpha \), with each entry \( \eta_j \) governing an independent observation \( y_j \),

\[ y_j \overset{\text{ind}}{\sim} g_{\eta_j}(\cdot) \quad \text{for } j = 1, 2, \ldots, n. \]  \hspace{1cm} (5.2)

All of this results in a \( p \)-parameter exponential family (3.1), with \( \alpha \) the canonical parameter vector. Letting \( \mu \) be the expectation vector of \( y = (y_1, \ldots, y_n)' \),

\[ \mu = E_{\alpha}\{y\}, \]  \hspace{1cm} (5.3)

the other entries of (3.1) are

\[ \hat{\beta} = X'y, \quad \beta = X'\mu, \quad \text{and} \quad \psi(\alpha) = \sum_{i=1}^{n} \phi(x_j\alpha) \]  \hspace{1cm} (5.4)

where \( x_j \) is the \( j \)th row of \( X \). The deviance difference \( \Delta(\beta) \) (3.11) has a simple form,

\[ \Delta(\beta) = (\alpha - \hat{\alpha})'(\beta + \hat{\beta}) - 2 \sum_{j=1}^{n} [\phi(x_j\alpha) - \phi(x_j\hat{\alpha})] \]

\[ = (\eta - \hat{\eta})'(\mu + \hat{\mu}) - 2 \sum_{j=1}^{n} [\phi(\eta_j) - \phi(\hat{\eta}_j)] \]  \hspace{1cm} (5.5)

(\( \hat{\alpha} \) the MLE of \( \alpha \), \( \hat{\eta} = X\hat{\alpha} \), and \( \hat{\mu} \) the expectation vector (5.3) corresponding to \( \alpha = \hat{\alpha} \)) according to (3.12).

As an extended example we now consider a microarray experiment discussed in Efron (2010), Section 2.1: 102 men, 50 healthy controls and 52 prostate cancer patients, have each had the activity of \( N = 6033 \) genes measured (Singh et al., 2002). A two-sample test comparing patients with controls has been performed for each gene, yielding a \( z \)-value \( z_k \), i.e., a test statistic having a standard normal distribution under \( H_{0k} \), the null hypothesis of no patient/control difference for gene \( k \),

\[ H_{0k} : z_k \sim N(0, 1). \]  \hspace{1cm} (5.6)

The experimenters, of course, are interested in identifying non-null genes.

Figure 5 shows a histogram of the \( N \) \( z \)-values. The standard normal curve is too high in the center and too low in the tails, suggesting that at least some of the genes are non-null. The better-fitting curve “Model 4” is a fit from the Poisson regression family discussed next.
We will assume that the $y_j$’s are independent Poisson observations, each having its own expectation $\mu_j$,

$$y_j \sim \text{Poi}(\mu_j) \quad j = 1, 2, \ldots, J,$$

(5.8)

and then fit curves to the histogram using Poisson regression. Why this might be appropriate is discussed at length in Efron (2008, 2010) but here we will just take it as a helpful example of the Bayes/bootstrap GLM modeling theory.

We consider Poisson regression models where the canonical parameters $\eta_j = \log(\mu_j)$ are $m$th-degree polynomial functions of the bin centers $x_j$, evaluated by $\text{glm}(y \sim \text{poly}(x,m),\text{Poisson})$ in the language R. This is a GLM with the Poisson family, $\eta_j = \log \mu_j$, where $X$ is a $J \times (m+1)$ matrix having rows $x_j = (1, x_j, x_j^2, \ldots, x_j^m)$ for $j = 1, 2, \ldots, J$. For the Poisson distribution, $\phi(\eta) = \mu$ in (5.1). The deviance difference function (5.5) becomes

$$\Delta(\beta) = (\eta - \hat{\eta})' (\mu + \hat{\mu}) - 2 \cdot 1' (\mu - \hat{\mu}),$$

(5.9)

1 a vector of $J$ ones.

Let “$M_m$” indicate the Poisson polynomial regression model of degree $m$. M2, with $\log(\mu_j)$ quadratic in $x_j$, amounts to a normal location-scale model for the marginal density of the $z_k$’s. Higher-order models are more flexible. M4, the quartic model, provided the heavy fitted curve in Figure 5. Table 2 shows the Poisson deviance for the fitted models M2 through M8. A dramatic decrease occurs between M3 and M4, but only slow change occurs after that. The AIC criterion for model $m$,

$$\text{AIC}(m) = \text{Deviance} + 2 \cdot (m + 1)$$

(5.10)
Table 2: Deviance from Poisson polynomial regression models for counts (5.7), prostate data; AIC criterion (5.10) is minimized for the quartic model M4. Boot % shows the proportion of each model selected in $B = 4000$ bootstrap replications of the AIC criterion, bootstrapping from M8. Bayes % are weighted Bayes proportions, assuming Jeffreys prior. The St Error column is obtained from the bootstrap-after-bootstrap calculations of Section 6.

<table>
<thead>
<tr>
<th>Model</th>
<th>Deviance</th>
<th>AIC</th>
<th>Boot %</th>
<th>Bayes %</th>
<th>(St Error)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M2</td>
<td>138.6</td>
<td>144.6</td>
<td>0%</td>
<td>0%</td>
<td>(0%)</td>
</tr>
<tr>
<td>M3</td>
<td>137.1</td>
<td>145.1</td>
<td>0%</td>
<td>0%</td>
<td>(0%)</td>
</tr>
<tr>
<td>M4</td>
<td>65.3</td>
<td>75.3</td>
<td>32%</td>
<td>36%</td>
<td>(20%)</td>
</tr>
<tr>
<td>M5</td>
<td>64.3</td>
<td>76.3</td>
<td>10%</td>
<td>12%</td>
<td>(14%)</td>
</tr>
<tr>
<td>M6</td>
<td>63.8</td>
<td>77.8</td>
<td>5%</td>
<td>5%</td>
<td>(8%)</td>
</tr>
<tr>
<td>M7</td>
<td>63.8</td>
<td>79.8</td>
<td>1%</td>
<td>2%</td>
<td>(6%)</td>
</tr>
<tr>
<td>M8</td>
<td>59.6</td>
<td>77.6</td>
<td>51%</td>
<td>45%</td>
<td>(27%)</td>
</tr>
</tbody>
</table>

is minimized at M4, though none of the subsequent models do much worse. The fit from M4 provided the “Model 4” curve in Figure 5.

Parametric bootstrap samples $y^*$ were generated from M4, as in (5.8),

$$y^*_j \overset{\text{ind}}{\sim} \text{Poi} (\hat{\mu}_j) \quad \text{for} \quad j = 1, 2, \ldots, J$$

with $\hat{\mu}_j$ the MLE values from M4. $B = 4000$ such samples were generated, and for each one the MLE $\hat{\alpha}$, and also $\hat{\beta}$ (5.4), were obtained from the R call glm($y^* \sim \text{poly}(x, 4), \text{poisson}$). Using the simplified notation $\alpha = \hat{\alpha}^*$ gives bootstrap vectors $\eta = X\alpha$, $\mu = \exp(\eta) = (\exp(\eta_j))$, $\beta = X'\mu$, where $X$ is the $49 \times 5$ matrix $\text{poly}(x, 4)$, and finally $\Delta(\beta)$ as in (5.9). (Notice that $\beta$ represents $\hat{\beta}$ here, not the “true value” $\beta$ of (5.4).)

The reweighted bootstrap distribution, with weights proportional to

$$w_i = e^{\Delta_i} \quad \text{on} \quad \beta_i \quad \text{for} \quad i = 1, 2, \ldots, B = 4000,$$

estimates the posterior distribution of $\beta$ given $\hat{\beta}$, starting from Jeffreys prior. The posterior expectation of any parameter $\theta = t(\beta)$ is estimated by $\sum w_it_i/\sum w_i$ as in (3.21).

We will focus attention on a false discovery rate (Fdr) parameter $\theta$,

$$\theta(z) = \text{Fdr}(z) = \left[ 1 - \Phi(z) \right] / \left[ 1 - F(z) \right]$$

where $\Phi$ is the standard normal cdf and $F(z)$ is the cdf of the Poisson regression model: in terms of the discretized situation (5.8),

$$F(z) = \sum_{x_j \leq z} \mu_j / \sum_{j=1}^{J} \mu_j$$

(with a “half count” correction at $x_j = z$). Fdr($z$) estimates the probability that a gene having $z_k$ exceeding the fixed value $z$ is non-null, as discussed for example in Efron (2008).

Figure 6 concerns the choice $z = 3$. Using quartic model M4 to estimate the $\mu_j$’s in (5.14) yields point estimate

$$\hat{\theta} = \text{Fdr}(3) = 0.192.$$
Figure 6: Posterior densities for $\theta = \text{Fdr}(3)$ (5.13), prostate data, based on $B = 4000$ parametric bootstrap replications (5.11) from the fourth-degree Poisson regression model M4. *Solid curve* Jeffreys Bayes posterior density, using (5.12); *heavy dashed curve* BCa confidence density (2.17). Both give 95% interval $\theta \in (-0.154, -0.241)$. *Light dashed curve* is unweighted bootstrap density. Total computation time was about 30 seconds.

Fdr values near 0.2 are in the “interesting” range where the gene might be reported as non-null, making it important to know the accuracy of (5.15).

The $B = 4000$ bootstrap samples for M4 (5.11) yield bootstrap replications $\theta_1, \theta_2, \ldots, \theta_B$. Their standard deviation is a bootstrap estimate of standard error for $\hat{\theta}$, $\hat{\text{se}} = 0.024$, so a typical empirical Bayes analysis might report $\text{Fdr}(3) = 0.0192 \pm 0.024$. A Jeffreys Bayes analysis gives the full posterior density of $\theta$ shown by the solid curve in Figure 6, with 95% credible interval

\[ \text{M4 : } \theta \in (0.154, 0.241). \tag{5.16} \]

In this case the BCa density (2.17) $((z_0, a) = (-0.047, -0.026))$ is nearly the same as the Bayes estimate, both of them lying just slightly to the left of the unweighted bootstrap density.

The choice of philosophy, Jeffreys Bayes or BCa frequentist, doesn’t make much difference here, but the choice of model does. Repeating the analysis using M8 instead of M4 to generate the bootstrap samples (5.11) sharply decreased the estimate. Figure 7 compares the bootstrap histograms; the 95% credible interval for Fdr(3) is now

\[ \text{M8 : } \theta \in (0.141, 0.239). \tag{5.17} \]

AIC calculations were carried out for each of the 4000 M8 bootstrap samples. Of these, 32% selected M4 as the minimizer, compared with 51% for M8, as shown in the *Boot %* column of Table 2. Weighting each sample proportionally to $\exp(\Delta_i)$ (5.12) narrowed the difference to 36% versus 45%, but still with a strong tendency toward M8.

It might be feared that M8 is simply justifying itself. However, standard nonparametric bootstrapping (resampling the $N_0 z_k$ values), gave slightly more extreme Boot percentages,

\[ 30\%(M4), 9\%(M5), 4\%(M6), 2\%(M7), 54\%(M8). \tag{5.18} \]

The fact is that data-based model selection is quite unstable here, as the accuracy calculations of Section 6 will verify.
6 Accuracy

Two aspects of our methodology’s Bayesian estimation accuracy are considered in this section: \textit{internal accuracy}, the bootstrap sampling error in estimates such as (3.7) (i.e., how many bootstrap replications \( B \) need we take?), and \textit{external accuracy}, statistical sampling error, for instance how much would the results in Figure 3 change for a new sample of 22 students? The i.i.d. (independent and identically distributed) nature of bootstrap sampling makes both questions easy to answer.

Internal accuracy is particularly straightforward, The estimate (3.7) for \( \hat{E}\{t(\beta)\|\hat{\beta}\} \) can be expressed in terms of \( s_i = t_i \pi_i R_i \) and \( r_i = \pi_i R_i \) as

\[
\hat{E} = \frac{\bar{s}}{\bar{r}} \quad \left( \bar{s} = \frac{1}{B} \sum_{1}^{B} s_i / B, \quad \bar{r} = \frac{1}{B} \sum_{1}^{B} r_i / B \right).
\]  

(6.1)

Let \( \overline{\text{cov}} \) be the \( 2 \times 2 \) empirical covariance matrix of the \( B \) vectors \((s_i, r_i)\). Then standard delta-method calculations yield a familiar approximation for the bootstrap coefficient of variation of \( \hat{E} \),

\[
\hat{cv}^2 = \frac{1}{B} \left( \frac{\bar{c}_{ss}}{\bar{s}^2} - 2 \frac{\bar{c}_{sr}}{\bar{s}\bar{r}} + \frac{\bar{c}_{rr}}{\bar{r}^2} \right)
\]

(6.2)

where \( \bar{c}_{ss}, \bar{c}_{sr}, \) and \( \bar{c}_{rr} \) are the elements of \( \overline{\text{cov}} \).

The Jeffreys Bayes estimate for eigenratio (4.11) was \( \hat{E} = 0.799 \) (nearly the same as the MLE 0.793). Formula (6.2) gave \( \hat{cv} = 0.002 \), indicating that \( \hat{E} \) nearly equaled the exact Bayes estimate \( E\{t(\beta)\|\hat{\beta}\} \). \( B = 10,000 \) was definitely excessive. Posterior parameters other than expectations are handled by other well-known delta-method approximations: the upper 97.5\% credible limit 0.241 in Figure 6 had \( \hat{cv} = 0.0015 \), again indicating near-zero bootstrap estimation error.

As far as external accuracy is concerned, the parametric bootstrap can be employed to assess its own sampling error, a “bootstrap-after-bootstrap” technique in the terminology of Efron.
(1992). Suppose we have calculated some Bayes posterior estimate \( \hat{Q} = Q(\hat{\beta}) \), for example \( \hat{E} \) or a credible limit, and wonder about its sampling standard error, i.e., its frequentist variability. As an answer, we sample \( K \) more times from \( f_{\hat{\beta}}(\cdot) \),

\[
f_{\hat{\beta}}(\cdot) \rightarrow \hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_K
\]

(6.3)

where the \( \gamma \) notation emphasizes that these replications are distinct from \( \beta_1, \beta_2, \ldots, \beta_B \) in (3.6), the original replications used to compute \( \hat{Q} \). Letting \( \hat{Q}_k = Q(\hat{\gamma}_k) \), the usual bootstrap estimate of standard error for \( \hat{Q} \) is

\[
\hat{se}(\hat{Q}) = \left[ \sum_{k=1}^{K} (\hat{Q}_k - \hat{Q})^2 / (K - 1) \right]^{1/2},
\]

(6.4)

\[\hat{Q} = \sum \hat{Q}_k / K. \quad K = 200 \] is usually plenty for reasonable estimation of \( se(\hat{Q}) \); see Table 6.2 of Efron and Tibshirani (1993).

This recipe looks arduous since each \( \hat{Q}_k \) requires \( B \) bootstrap replications for its evaluation. Happily, a simple reweighting scheme on the original \( B \) replications finesses all that computation. Define

\[
W_{ki} = f_{\beta_i}(\hat{\gamma}_k) / f_{\beta_i}(\hat{\beta}).
\]

(6.5)

**Lemma 4.** If \( \hat{Q} \) is a posterior expectation \( \hat{E} = \sum t_i \pi_i R_i / \sum \pi_i R_i \), then the importance sampling estimate of \( \hat{Q}_k \) is

\[
\hat{Q}_k = \sum_{i=1}^{B} t_i \pi_i R_i W_{ki} / \sum_{i=1}^{B} \pi_i R_i W_{ki};
\]

(6.6)

for general quantities \( \hat{Q} \), reweighting \( \beta_i \) proportionately to \( \pi_i R_i W_i \) gives \( \hat{Q}_k \).

The proof of Lemma 4 follows immediately from

\[
R_i W_{ki} = f_{\beta_i}(\hat{\gamma}_k) / f_{\beta_i}(\hat{\beta}),
\]

(6.7)

which is the correct importance sampling factor for converting an \( f_{\beta}(\beta) \) sample into an \( f_{\beta}(\hat{\gamma}_k) \) likelihood.

Formula (6.6) requires no new computations of \( t(\beta) \), \( \pi(\beta) \), or \( R(\beta) \), and in exponential families the factor \( W_{ki} \) is easily calculated:

\[
W_{ki} = e^{(\alpha_i - \hat{\alpha})(\hat{\gamma}_k - \hat{\beta})}
\]

(6.8)

where \( \alpha_i \) is the canonical vector in (3.1) corresponding to \( \beta_i \). This usually makes the computation for the bootstrap-after-bootstrap standard error (6.4) much less than that needed originally for \( \hat{Q} \). (Note: formula (6.5) is invariant under smooth transformations of \( \beta \), and so \( W_{ki} \) can be calculated directly in other coordinate systems as a ratio of densities.)

Returning to the 97.5% credible limit 0.241 of Figure 6, \( K = 400 \) bootstrap-after-bootstrap replications in (6.5) gave estimated standard error (6.4),

\[
\hat{se} = 0.030.
\]

(6.9)

This brings up an important inferential distinction concerning the value 0.241: from a Bayesian viewpoint, it is (nearly) exact as the 97.5% credible limit given \( \hat{\beta} \), at least if we take Jeffreys
prior literally. However we can see from (6.9) that it has substantial frequentist sampling error and of course in this case there is no real reason to take the prior literally.

A more striking use of (6.4) appears in the last column of Table 2, Section 5. Let \( t_4(\beta_i) \) be the indicator function of whether or not Model 4 minimized AIC for the \( i \)th bootstrap replication: \( \hat{E}\{t_4(\beta) | \beta \} = 0.36 \) according to the Bayes \% column. However, its bootstrap-after-bootstrap standard error estimate was \( \hat{se} = 0.20 \), with similarly enormous standard errors for the other model selection probabilities. From a frequentist viewpoint, data-based model selection will be a highly uncertain enterprise here.

Frequentist assessment of objective Bayes procedures has been advocated in the literature, for example in Berger (2006) and Gelman et al. (1996), but seems to be followed most often in the breach. The methodology here can be useful for injecting a note of frequentist caution into Bayesian data analysis based on priors of convenience.

If our original data set \( y \) consists of \( n \) i.i.d. vectors \( y_i \), as in Table 1, we can jackknife instead of bootstrapping the \( \hat{\gamma}_k \)’s. Now \( \hat{\gamma}_k \) is \( \hat{\beta} \) recomputed from the data set \( y_{(i)} \) having \( y_i \) removed for \( k = 1, 2, \ldots, n \). Formulas (6.5)–(6.8) still hold, yielding

\[
\hat{se}_{\text{jack}} = \left[ \frac{n-1}{n} \sum_{k=1}^{n} \left( \hat{Q}_k - \bar{Q} \right)^2 \right]^{1/2}.
\]

An advantage of jackknife resampling is that the \( \hat{\gamma}_k \) values lie closer to \( \hat{\beta} \), making \( W_{ki} \) closer to 1 and putting less strain on the importance sampling formula (6.6).

Variations in the parametric bootstrap methodology are possible in our original calculations as well as here. Instead of (3.6), for instance, we might obtain \( \beta_1, \beta_2, \ldots, \beta_B \) from \( N_p(\hat{\mu}_\beta, \hat{\Sigma}_\beta) \), where \( \hat{\mu}_\beta \) and \( \hat{\Sigma}_\beta \) are the observed mean and covariance of \( \beta \)’s from a preliminary \( f(\hat{\beta}) \) sample. This changes the conversion factor \( R(\hat{\beta}) \) (3.8), but in an easily computable way. Or we might try resampling from \( N_p(\hat{\mu}_\beta, c\hat{\Sigma}_\beta) \) with \( c > 1 \), in order to tame extreme values of \( R(\hat{\beta}) \). The internal error formula (6.2) always provides a useful check on the adequacy of importance sampling calculations.

7 Summary

The main points made by the theory and examples of the preceding sections are as follows:

- The parametric bootstrap distribution is a favorable starting point for importance sampling computation of Bayes posterior distributions (as in Figure 2).
- This computation is implemented by reweighting the bootstrap replications rather than by drawing observations directly from the posterior distribution as with MCMC (formulas (3.7), (3.8)).
- The necessary weights are easily computed in exponential families for any prior, but are particularly simple starting from Jeffreys invariant prior, in which case they depend only on the deviance difference \( \Delta(\beta) \) ((3.9)–(3.12), (3.21), (4.7), (5.5)).
- The deviance difference depends asymptotically on the skewness of the family, having a cubic normal form (3.29).
- In our examples, Jeffreys prior yielded posterior distributions not much different than the unweighted bootstrap distribution. This may be unsatisfactory for single parameters of interest in multiparameter families (Figure 3).
• Better uninformative priors, such as the Welch–Peers family or reference priors, are closely related to the frequentist BCa reweighting formula ((2.17), Figure 2, Figure 6).

• Because of the i.i.d. nature of bootstrap resampling, simple formulas exist for the accuracy of posterior computations as a function of the number $B$ of bootstrap replications. Even with excessive choices of $B$, computation time was measured in seconds for our examples (6.2).

• An efficient second-level bootstrap algorithm (“bootstrap-after-bootstrap”) provides estimates for the frequentist accuracy of Bayesian inferences ((6.3)–(6.6)).

• This can be important in assessing inferences based on formulaic priors, such as those of Jeffreys, rather than on genuine prior experience (last column, Table 2 of Section 5).

References


### A Appendix

*Transformation of coordinates:* Let \( J(\beta) \) be the Jacobian of the transformation \( \gamma = m(\beta) \), that is, the absolute value of the determinant of the Hessian matrix \( (\partial^2 \hat{\gamma}_i / \partial \gamma_j) \). Then \( \tilde{f}_\gamma(\hat{\gamma}) = f_\beta(\hat{\beta})J(\hat{\beta}) \) gives

\[
\tilde{\xi}(\gamma) = \frac{\tilde{f}_\gamma(\hat{\gamma})J(\hat{\beta})}{\tilde{f}_\xi(\gamma)J(\beta)} = \xi(\beta)\frac{J(\hat{\beta})}{J(\beta)}
\]

in (4.2), and

\[
\tilde{R}(\gamma) = \frac{\tilde{f}_\gamma(\hat{\gamma})}{\tilde{f}_\xi(\gamma)} = \frac{f_\beta(\hat{\beta})J(\hat{\beta})}{f_\beta(\beta)J(\beta)} = \frac{R(\beta)J(\hat{\beta})}{J(\beta)} = \xi(\beta)e^{\Delta(\beta)}\frac{J(\hat{\beta})}{J(\beta)} = \tilde{\xi}(\gamma)e^{\tilde{\Delta}(\gamma)},
\]

since \( \tilde{\Delta}(\gamma) = \Delta(\beta) \) by the transformation invariance of the deviance.
For any prior density \( \pi(\beta) \) we have \( \tilde{\pi}(\gamma) = \pi(\beta)J(\beta) \) and
\[
\tilde{\pi}(\gamma)\tilde{R}(\gamma) = \pi(\beta)J(\beta)R(\beta)J(\hat{\beta})/J(\beta)
\]
\[
= J(\hat{\beta})\pi(\beta)R(\beta). \tag{A.3}
\]
\( J(\hat{\beta}) \) acts as a constant in (A.3), showing that (4.4) is identical to (3.7). This also applies to Jeffreys prior, \( \tilde{\pi}\text{Jeff}(\gamma) \), which by design is transformation invariant, yielding (4.5).

**Theorem 1:** In a one-parameter exponential family, (3.13)–(3.14) give
\[
\psi(\alpha) - \psi(\hat{\alpha}) \equiv \hat{\beta}d\alpha + \hat{V}(d\alpha)^2/2 + \hat{U}(d\alpha)^3/6 \tag{A.4}
\]
and
\[
\beta - \hat{\beta} \equiv \hat{V}d\alpha + \hat{U}(d\alpha)^2/2 \tag{A.5}
\]
where \( d\alpha = \alpha - \hat{\alpha} \), \( \hat{V} = V(\hat{\alpha}) \), and \( \hat{U} = U(\hat{\alpha}) \). Expression (3.12) for \( \Delta \) can be written as
\[
\Delta = \left( \beta - \hat{\beta} \right) d\alpha + 2 \left[ \hat{\beta}d\alpha - (\psi - \hat{\psi}) \right]. \tag{A.6}
\]
Applying (A.4)–(A.5) reduces (A.6) to
\[
\Delta \equiv \frac{1}{6} \hat{U}(d\alpha)^3 = \frac{1}{6} \hat{\gamma} \left[ \hat{V}^{\frac{1}{2}} (\alpha - \hat{\alpha}) \right]^3 = \frac{1}{6} \hat{\gamma}Z^3 \tag{A.7}
\]
with \( \hat{\gamma} = \hat{U}/\hat{V}^{3/2} \) the skewness, the last line following from \( Z \equiv \hat{V}^{-1/2}(\beta - \hat{\beta}) \equiv \hat{V}^{1/2}(\alpha - \hat{\alpha}) \) (A.5). Standard exponential family theory shows that \( Z \rightarrow \mathcal{N}(0,1) \) under repeated sampling, verifying the theorem (remembering that the asymptotics here are for \( \beta \sim f_\beta(\cdot) \), with \( \hat{\beta} \) fixed).

The skewness \( \hat{\gamma} \) is then \( O(n^{-1/2}) \), making \( \Delta \) of order \( O_p(n^{-1/2}) \). The first missing term in the Taylor expansion (A.7) for \( \Delta \) is \( \hat{\delta}Z^4/12 \), \( \hat{\delta} \) the kurtosis, and is of order \( O_p(n^{-1}) \).

The multiparameter version of Theorem 1 begins by considering a one-parameter subfamily of (3.1) now indexed by \( \alpha \) rather than \( \beta \),
\[
f_a^{(v)}(\hat{\beta}) = f_{\hat{\alpha}+av}(\hat{\beta}) = e^{(\hat{\alpha}+av)\beta - \psi(\hat{\alpha}+av)}f_0(\hat{\beta}) \tag{A.8}
\]
where \( v \) is some fixed vector in \( \mathbb{R}^p \); \( a \) here is not connected with that in (2.17). The deviance difference within \( f_a^{(v)}(\cdot) \) is
\[
\Delta^{(v)}(a) = \Delta(\hat{\alpha} + av) \tag{A.9}
\]
since deviance is entirely determined by the two densities involved.

The exponential family terms (3.1) for family \( f_a^{(v)}(\cdot) \) are
\[
\alpha^{(v)} = a, \quad \beta^{(v)} = v'\hat{\beta}, \quad \beta^{(v)} = v'\beta, \quad \hat{V}^{(v)} = v'\hat{V}v,
\]
and
\[
\hat{U}^{(v)} = \sum_{j=1}^{p} \sum_{k=1}^{p} \sum_{l=1}^{p} \hat{U}_{jkl}v_jv_kv_l, \tag{A.10}
\]
giving skewness \( \hat{\gamma}^{(v)} = \hat{U}^{(v)}/\hat{V}^{(v)3/2} \). Applying the one-dimensional result gives
\[
\Delta(\hat{\alpha} + av) \equiv \frac{1}{6} \hat{\gamma}^{(v)}Z^{(v)3} \quad \text{with} \quad Z^{(v)} = \frac{(\beta - \hat{\beta})}{(v'\hat{V}v)^{1/2}} \tag{A.11}
\]
Since \( v \) can be any vector, (A.11) describes the asymptotic form of \( \Delta(\cdot) \) in the neighborhood of \( \hat{\alpha} \).
Theorem 2: For a single observation \( y \sim \mathcal{N}_d(\mu, \Sigma) \), let \( f_1 \) and \( f_2 \) represent its density under \((\mu_1, \Sigma_1)\) and \((\mu_2, \Sigma_2)\), respectively. Then

\[
2 \log \frac{f_1(y)}{f_2(y)} = \log |\Sigma_1| + (y - \mu_2)'\Sigma_2(y - \mu_2) - (y - \mu_1)'\Sigma_1(y - \mu_1).
\] (A.12)

But if \( y \sim \mathcal{N}_d(\mu_1, \Sigma_1) \),

\[
E_{f_1} \left\{ (y - \mu_2)'\Sigma_2^{-1}(y - \mu_2) \right\} = (\mu_2 - \mu_1)'\Sigma_2^{-1}(\mu_2 - \mu_1) + \text{tr} \Sigma_1 \Sigma_2^{-1}
\] (A.13)

while \( E_{f_1} \left\{ (y - \mu_1)'\Sigma_1^{-1}(y - \mu_1) \right\} = d \). Taking the \( f_1 \) expectation of (A.12) gives the deviance

\[
D((\mu_1, \Sigma_1), (\mu_2, \Sigma_2)) = \log |\Sigma_2|/|\Sigma_1| + (\mu_2 - \mu_1)'\Sigma_2^{-1}(\mu_2 - \mu_1) + \text{tr} \Sigma_1 \Sigma_2^{-1} - d
\] (A.14)

for sample size \( n = 1 \). The deviance difference for sample size \( n \)

\[
\Delta = \frac{n}{2} \left\{ D((\mu, \Sigma)), (\hat{\mu}, \hat{\Sigma}) \right\} - D((\hat{\mu}, \hat{\Sigma}), (\mu, \Sigma))
\] (A.15)

is then seen to equal (4.7).

The density of \((\hat{\mu}, \hat{\Sigma})\) from a \( \mathcal{N}_p(\mu, \Sigma) \) sample of size \( n \) is proportional to

\[
\left\{ |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2} (\mu - \hat{\mu})' \Sigma^{-1} (\mu - \hat{\mu})} \right\} \left\{ \hat{\Sigma}^{\frac{n-d-2}{2}} e^{-\frac{1}{2} \text{tr} \Sigma^{-1} \hat{\Sigma}} / |\Sigma|^{\frac{n-1}{2}} \right\},
\] (A.16)

yielding (4.6).

**Table 3: BCa constants \( z_0 \) and \( a \) for our three examples.**

<table>
<thead>
<tr>
<th></th>
<th>( \hat{\theta} )</th>
<th>( z_0 )</th>
<th>( a )</th>
</tr>
</thead>
<tbody>
<tr>
<td>student correlation</td>
<td>.498</td>
<td>-.069</td>
<td>0</td>
</tr>
<tr>
<td>student eigenratio</td>
<td>.793</td>
<td>-.222</td>
<td>0</td>
</tr>
<tr>
<td>prostate data Fdr(3)</td>
<td>.192</td>
<td>-.047</td>
<td>-.026</td>
</tr>
</tbody>
</table>

The BCa weights: The BCa system of second-order accurate bootstrap confidence intervals was introduced in Efron (1987), and restated in weighting form (2.17) in Efron and Tibshirani (1998). The bias correction constant \( z_0 \) is obtained directly from the MLE \( \hat{\theta} \) and the bootstrap replication \( \theta_1, \theta_2, \ldots, \theta_B \) according to

\[
z_0 = \Phi^{-1} \left( \frac{\# \left\{ \theta_i \leq \hat{\theta} \right\}}{B} \right).
\] (A.17)

DiCiccio and Efron’s (1992) paper presents the “ABC” algorithm for compacting \( a \), the acceleration constant, but this requires special functions for each exponential family. A more computer-intensive R program for \( a \)'s calculation, `accel`, is available from the author.

Table 3 shows \( z_0 \) and \( a \) for our three main examples. Notice the especially large bias correction needed for the eigenratio.