




## The Automatic Construction of Bootstrap Confidence Intervals

Bradley Efron & Balasubramanian Narasimhan


To cite this article: Bradley Efron & Balasubramanian Narasimhan (2020): The Automatic Construction of Bootstrap Confidence Intervals, Journal of Computational and Graphical Statistics, DOI: [10.1080/10618600.2020.1714633](https://doi.org/10.1080/10618600.2020.1714633)

To link to this article: <https://doi.org/10.1080/10618600.2020.1714633>

 View supplementary material 

 Accepted author version posted online: 14 Jan 2020.  
Published online: 12 Mar 2020.

 Submit your article to this journal 

 Article views: 132

 View related articles 

 View Crossmark data 



# The Automatic Construction of Bootstrap Confidence Intervals

Bradley Efron<sup>a,b</sup> and Balasubramanian Narasimhan<sup>a</sup>

<sup>a</sup>Department of Statistics, Stanford University, Stanford, CA; <sup>b</sup>Department of Biomedical Data Science, Stanford University, Stanford, CA

## ABSTRACT

The standard intervals, for example,  $\hat{\theta} \pm 1.96\hat{\sigma}$  for nominal 95% two-sided coverage, are familiar and easy to use, but can be of dubious accuracy in regular practice. Bootstrap confidence intervals offer an order of magnitude improvement—from first order to second order accuracy. This article introduces a new set of algorithms that automate the construction of bootstrap intervals, substituting computer power for the need to individually program particular applications. The algorithms are described in terms of the underlying theory that motivates them, along with examples of their application. They are implemented in the R package `bcaboot`. Supplementary materials for this article are available online.

## ARTICLE HISTORY

Received November 2018  
Revised December 2019

## KEYWORDS

bca method; Exponential families; Nonparametric intervals; Second-order accuracy

## 1. Introduction

Among the most useful, and most used, of statistical constructions are the *standard intervals*

$$\hat{\theta} \pm z^{(\alpha)}\hat{\sigma}, \quad (1.1)$$

giving approximate confidence statements for a parameter  $\theta$  of interest. Here  $\hat{\theta}$  is a point estimate,  $\hat{\sigma}$  an estimate of its standard error, and  $z^{(\alpha)}$  the  $\alpha$ th quantile of a standard normal distribution. If  $\alpha = 0.975$ , for instance, interval (1.1) has approximate 95% coverage.

A prime virtue of the standard intervals is their ease of application. In principle, a single program can be written that automatically produces intervals (1.1), say with  $\hat{\theta}$  a maximum likelihood estimate (MLE) and  $\hat{\sigma}$  its bootstrap standard error. Accuracy, however, can be a concern: if  $X \sim \text{Poisson}(\theta)$  is observed to be 16, then (1.1) gives 95% interval (8.16, 23.84) for  $\theta$ , compared to the exact interval (9.47, 25.41) (obtained via the usual Neyman construction). Modern theory and computer power now provide an opportunity for the routine use of more accurate intervals, this being the theme of what follows.

Exact intervals for  $\theta$  are almost never available in the presence of nuisance parameters, which is to say, in most applied settings. This accounts for the popularity of the standard intervals. Bootstrap confidence intervals (Efron 1987; DiCiccio and Efron 1996) were proposed as general-purpose improvements over (1.1). They are *second-order accurate*, having the error in their claimed coverage probabilities going to zero at rate  $O(1/n)$  in sample size compared to the first-order rate  $O(1/\sqrt{n})$  for standard intervals. This is often a


substantial improvement, yielding bootstrap limits (9.42, 25.53) compared to the exact limits (9.47, 25.41) in the Poisson problem.<sup>1</sup>

Bootstrap standard error estimates have been used an enormous number of times, perhaps in the millions, but the story is less happy for bootstrap confidence intervals. Various limitations of previous bootstrap packages such as `bootstrap` have discouraged everyday use. The new suite of programs introduced here, `bcaboot`, has been designed to overcome four impediments to the routine “automatic” application of second-order accurate bootstrap confidence intervals:

1. For nonparametric intervals, in problems with  $n$  original observations, previous programs required  $n$  extra recomputations of the statistic of interest (for the assessment of the acceleration  $a$ ) in addition to the  $B \sim 2000$  replications needed to estimate the bootstrap histogram. This becomes a burden for contemporary sample sizes of say  $n = 10,000$ . The new algorithm `bcajack` introduces a “folding” mechanism that groups the observations, say into 50 groups of size 200 each, now with  $n$  effectively only 50. This is done without requiring any extra intervention from the user. An alternate nonparametric confidence interval algorithm `bcajack2` employs a novel theoretical approach to calculate the acceleration  $a$  without any need for the extra  $n$  recomputations
2. Parametric models, such as multivariate normal or logistic regression, are where bootstrap intervals show themselves to best advantage, but previous packages have required case-by-case special forms from the statistician. The new algorithm `bcapar` allows the user to calculate bootstrap replications using whatever function gave the original estimate,

**CONTACT** Bradley Efron  [brad@stat.stanford.edu](mailto:brad@stat.stanford.edu)  Department of Statistics, Stanford University, Stanford, CA 94305.

<sup>1</sup>Using the bca algorithm (2.2). The Neyman exact intervals were calculated by splitting the atom of Poisson probability at 16 in half. Following the same convention, the bootstrap limits (9.42, 25.53) gave exceedance Poisson probabilities 0.0240 and 0.0238, compared with the nominal value 0.0250. The standard interval exceedances were 0.007 and 0.048, showing that the interval norms were too long on the left and too short on the right. Notice that even though the Poisson distribution is discrete, the Poisson parameter  $\theta$  varies continuously, allowing the second-order accuracy theory to apply.

 Supplementary materials for this article are available online. Please go to [www.tandfonline.com/r/JCGS](http://www.tandfonline.com/r/JCGS).

irregardless of its form. Again, this has to do with a novel characterization of the acceleration  $a$ .

3. Both `bcajack` and `bcapar` allow the user to separately calculate the bootstrap replications. This is advantageous for large-scale applications where each replication takes substantial time, and where distributed computation may be necessary.
4. `bcajack` and `bcapar` provide estimates of the interval endpoints Monte Carlo errors, helping guide the choice of the number of bootstrap replications  $B$  necessary for reasonable accuracy. These are based on a novel use of jackknife theory, which requires almost no increase in computation.

Figure 1 concerns an application of `bcajack`. The diabetes data, Table 1 of Efron (2004), comprises  $n = 442$  vectors in  $\mathcal{R}^{11}$ ,

$$v_i = (x_i, y_i), \quad i = 1, 2, \dots, 442; \quad (1.2)$$

$x_i$  is a 10-vector of baseline predictors—age, sex, BMI, etc.—for patient  $i$ , while  $y_i$  is a real-valued measure of disease progression one year later. The data consist of

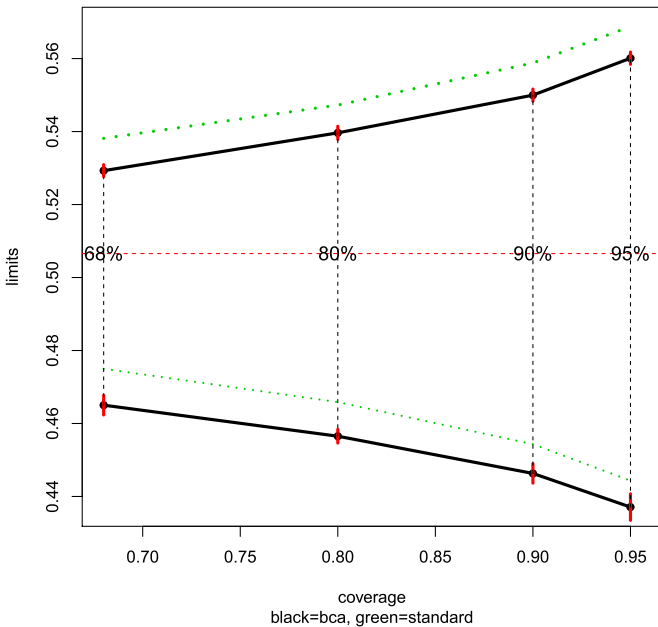
$$X \text{ and } y, \quad (1.3)$$

where  $X$  is the  $n \times 10$  matrix with  $i$ th row  $x_i$ , and  $y$  the  $n$ -vector of  $y_i$  values. The full dataset can be expressed as an  $n \times 11$  matrix  $v$ , having  $v_i$  as the  $i$ th row.

Suppose we are interested in the adjusted  $R^2$  for ordinary linear regression of  $y$  on  $X$ , with point estimate  $\hat{\theta} = 0.507$  for the diabetes data,

$$\hat{R}_{\text{adj}}^2 = \hat{R}^2 - (1 - \hat{R}^2)p/(n - p - 1) \quad (1.4)$$

with  $n = 442$  and  $p = 10$  here. How accurate is this estimate?



**Figure 1.** Two-sided nonparametric confidence limits for adjusted  $R^2$ , diabetes data, plotted vertically versus nominal coverage level. Bootstrap (solid curves); standard intervals (dotted curves). Small red vertical bars indicate Monte Carlo error in the bootstrap curves, from program `bcajack` (Section 3). Horizontal dashed line shows  $\hat{\theta} = 0.507$ , seen to lie closer to upper bootstrap limits than to lower ones.

The heavy curves in Figure 1 describe the lower and upper limits of nonparametric bootstrap confidence intervals for the true  $\theta$ , with nominal two-sided coverages ranging from 68% at the left of the figure to 95% at the right. At 68%, the bootstrap limits are

$$(0.462, 0.529), \quad (1.5)$$

compared to the standard limits (dotted curves)

$$(0.475, 0.538), \quad (1.6)$$

from (1.1) with  $\alpha = 0.84$ . The bootstrap standard error for  $\hat{\theta}$  was  $\hat{\sigma} = 0.032$ , so the bootstrap limits are shifted downward from the standard intervals by somewhat less than half a standard error. Figure 1 says that  $\hat{\theta}$  is substantially biased upward for the adjusted  $R^2$ , requiring downward corrections for second-order accuracy. (Bias corrections for  $\hat{\theta}$  are briefly discussed near the end of Section 3.)

The calculations for Figure 1 came from the nonparametric bootstrap confidence interval program `bcajack` discussed in Section 3. The call was

$$\text{bcajack}(v, 2000, \text{radj}), \quad (1.7)$$

with  $v$  the diabetes data matrix,  $B = 2000$  the number of bootstrap replications, and `radj` a program for  $\hat{\theta}$ : for any matrix  $v^*$  having 11 columns `radj`( $v^*$ ) linearly regresses the 11th column on the first 10 and returns the adjusted  $R^2$ .

There is nothing special about adjusted  $R^2$ . For any parameter of interest, say  $\hat{\phi} = r(v)$ , the call `bcajack`( $v, 2000, r$ ) produces the equivalent of Figure 1. All effort has been transferred from the statistician to the computer. It is in this sense that `bcajack` deserves to be called “automatic.”

Was the number of bootstrap replications,  $B = 2000$ , too many or perhaps too few? `bcajack` also provides a measure of *internal error*, an estimate of standard deviation of the Monte Carlo bootstrap confidence limits due to stopping at  $B$  replications. The vertical red bars in Figure 1 indicate  $\pm$  one such internal standard deviation. In this example,  $B = 2000$  was sufficient and not greatly wasteful. Stopping at  $B = 1000$  would magnify the Monte Carlo error by about  $\sqrt{2}$ , perhaps becoming uncomfortably large at the lower bootstrap confidence limits.

*Parametric* bootstrap confidence intervals are inherently more challenging to automate. All nonparametric problems have the same basic structure, whereas parametric models differ from each other in their choice of sufficient statistics. Working within this limitation, the program `bcapar` (Section 4) minimizes the required input from the statistician: bootstrap replications of the statistic of interest along with their bootstrap sufficient statistics. This is particularly convenient for generalized linear models, where the sufficient vectors are immediately available.

One can ask whether second-order corrections are worthwhile in ordinary situations. The answer depends on the importance attached to the value of the parameter of interest  $\theta$ . As a point of comparison, the usual Student’s  $t$  intervals  $\bar{x} \pm t_n^{(\alpha)} \hat{\sigma}$  make corrections of order only  $O(\hat{\sigma}/n)$  to (1.1), while the bootstrap corrections are  $O(\hat{\sigma}/\sqrt{n})$ . An obvious comment, but a crucial one, is that modern computing power makes  $B = 2000$

bootstrap replications routinely feasible in a great range of real applications. (The calculations for Figure 1 took 3 sec.)

We proceed as follows: Section 2 gives a brief review of bootstrap confidence intervals. Section 3 discusses nonparametric intervals and their implementation in `bcajack`. Parametric bootstrap intervals, and their implementation by the program `bcapar`, are discussed in Section 4. Comments on the programs appear in the Appendix.

Barndorff-Nielsen (1983) initiated the study of advanced likelihood methods for the construction of second-order (and higher) accurate confidence intervals (see, e.g., Barndorff-Nielsen and Cox 1994). Skovgaard (1985) and several others carried on this line of work. The result has been an elegant asymptotic theory, but one that has proved difficult to apply in practice. More recently, Pierce and Bellio (2017) have developed a more practical version of likelihood confidence intervals, based on deviance residuals in exponential families, as is the work of DiCiccio and Young (2008). An example of the former is given in Section 4.

The bca bootstrap approach, featured here, is less elegant than the likelihood theory, but easier to use. As the *automatic* in our title implies, the new bca algorithms have been developed with ease of application in mind. The basic bca theory is reviewed in what follows, without rigorous justification. DiCiccio and Efron (1996) give more of the background, while Hall (1992) is the standard theoretical reference.

## 2. Bootstrap Confidence Intervals

We present a brief review of the bca theory for bootstrap confidence intervals, which underlies algorithms `bcajack` and `bcapar`. More thorough expositions appear in DiCiccio and Efron (1996) and Chapter 11 of Efron and Hastie (2016).

The standard method level- $\alpha$  endpoint

$$\hat{\theta}_{\text{stan}}(\alpha) = \hat{\theta} + z^{(\alpha)}\hat{\sigma} \quad (2.1)$$

depends on two summary statistics,  $\hat{\theta}$  and  $\hat{\sigma}$ . The bca level- $\alpha$  endpoints (Efron 1987),

$$\hat{\theta}_{\text{bca}}(\alpha) = G^{-1}\Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha)}}{1 - \hat{a}(\hat{z}_0 + z^{(\alpha)})}\right) \quad (2.2)$$

as plotted in Figure 1 ( $\Phi$  the standard normal cdf), require three other quantities:  $G(\cdot)$ , the cdf of the bootstrap distribution of  $\hat{\theta}$ , and two auxiliary constants discussed below: the *bias corrector*  $\hat{z}_0$  and the *acceleration*  $\hat{a}$ . In almost all cases,  $G$ ,  $\hat{z}_0$ , and  $\hat{a}$  must be obtained from Monte Carlo simulations. Doing the bca calculations in fully automatic fashion is the goal of the programs presented in this article.

The standard intervals (1.1) take literally the first-order approximation

$$\hat{\theta} \sim \mathcal{N}(\theta, \sigma^2), \quad (2.3)$$

which can be quite inaccurate in practice. A classical tactic is to make a monotone transformation  $m(\theta)$  to a more favorable scale,

$$\hat{\phi} = m(\hat{\theta}) \quad \text{and} \quad \phi = m(\theta), \quad (2.4)$$

compute the standard interval on the  $\phi$  scale, and map the endpoints back to the  $\theta$  scale by the inverse transformation  $m^{-1}(\cdot)$ . In the classical example,  $m(\cdot)$  is Fisher's  $z$ -transformation for the normal correlation coefficient. For the Poisson family  $\hat{\theta} \sim \text{Poisson}(\theta)$ , the square root transformation  $m(\theta) = \theta^{1/2}$  is often suggested.

This argument assumes the existence of a transformation (2.4) to a normal translation family,

$$\hat{\phi} \sim \mathcal{N}(\phi, 1). \quad (2.5)$$

(If the variance of  $\hat{\phi}$  is any constant it can be reduced to 1 by rescaling.) The bca formula (2.2) is motivated by a less restrictive assumption: that there exists a monotone transformation (2.4) such that

$$\hat{\phi} \sim \mathcal{N}(\phi - z_0\sigma_\phi, \sigma_\phi^2), \quad \sigma_\phi = 1 + a\phi. \quad (2.6)$$

Here the *bias corrector*  $z_0$  allows for a bias in  $\hat{\phi}$  as an estimate of  $\phi$ , while the *acceleration*  $a$  allows for nonconstant variance. Lemma 1 of Efron (1987) shows that (2.4)–(2.6) implies that the bca formula (2.2) gives correct confidence interval endpoints in a strong sense of correctness: the bca method is motivated by a (hidden) transformation to a simple translation model in which the exact interval endpoints are obvious, these transforming back to formula (2.2). Section 8 of DiCiccio and Efron (1992) gives a rigorous discussion.

Crucially, *knowledge of the transformation (2.4) is not required*; only its existence is assumed, serving as motivation for the bca intervals. Formula (2.2) is *transformation invariant*: for any monotone transformation (2.4), the bca endpoints transform correctly,

$$\hat{\phi}_{\text{bca}}(\alpha) = m(\hat{\theta}_{\text{bca}}(\alpha)). \quad (2.7)$$

This is not true for the standard intervals, making them vulnerable to poor performance if applied on the “wrong” scale.

Under reasonably general conditions, Hall (1988) and others showed that formula (2.2) produces second-order accurate confidence intervals. See also the derivation in Section 8 of DiCiccio and Efron (1996).

Formula (2.2) makes three corrections to (2.1):

1. for nonnormality of  $\hat{\theta}$  (through the bootstrap cdf  $G$ );
2. for bias of  $\hat{\theta}$  (through  $\hat{z}_0$ );
3. for nonconstant standard error of  $\hat{\theta}$  (through  $\hat{a}$ ).

In various circumstances, all three can have substantial effects.

The parameters  $z_0$  and  $a$ , and their estimates  $\hat{z}_0$  and  $\hat{a}$ , are interesting in their own right. Standard intervals (1.1) depend on two estimated parameters,  $\hat{\theta}$  and  $\hat{\sigma}$ ; the approximate bca intervals called “abc” in Section 4 require just three more— $\hat{z}_0$ ,  $\hat{a}$ , and a measure of skewness of the bootstrap distribution  $\mathcal{G}$ —to achieve second-order accuracy; see the discussion in Sections 3 and 4 of DiCiccio and Efron (1996). This simplifies the theory of second-order confidence intervals, and underlies the ease of application of `bcajack` and its parametric counterpart `bcapar` (Section 4).

As a particularly simple parametric example, where all the calculations can be done theoretically rather than by simulation, suppose we observe

$$\hat{\theta} \sim \theta \cdot \text{Gamma}_{10/10}, \quad (2.8)$$

$\text{Gamma}_{10}$  a gamma variable with 10 degrees of freedom (density  $\theta^9 \exp(-\theta) / \Gamma(10)$ ,  $\theta > 0$ ) so  $\theta$  is the expectation of  $\hat{\theta}$ .

Table 1 shows confidence limits at  $\alpha = 0.025, 0.16, 0.84$ , and  $0.975$ , having observed  $\hat{\theta} = 1$ . (Any other value of  $\hat{\theta}$  simply multiplies the limits.) Five different methods are evaluated: (1) the standard endpoint (2.1); (2) the percentile method  $\hat{\theta}_{\text{pct}}(\alpha) = G^{-1}(\alpha)$ , that is, taking  $z_0$  and  $a = 0$  in (2.2); (3) the bias-corrected method  $\hat{\theta}_{\text{bc}}(\alpha)$ , setting  $a = 0$  in (2.2); (4) the full bca endpoint (2.2); (5) the exact endpoint. Accuracy of the approximations increases at each step, culminating in near three-digit accuracy for the full bca method.

The three elements of the bca formula— $G$ ,  $z_0$ , and  $a$ —must usually be estimated by simulation. Figure 2 shows the histogram of 2000 nonparametric bootstrap replications of  $\hat{\theta}$ , the adjusted  $R^2$  statistic for the diabetes data example of Figure 1. Its cdf  $\hat{G}$  estimates  $G$  in (2.2).

We also get an immediate estimate for  $z_0$ ,

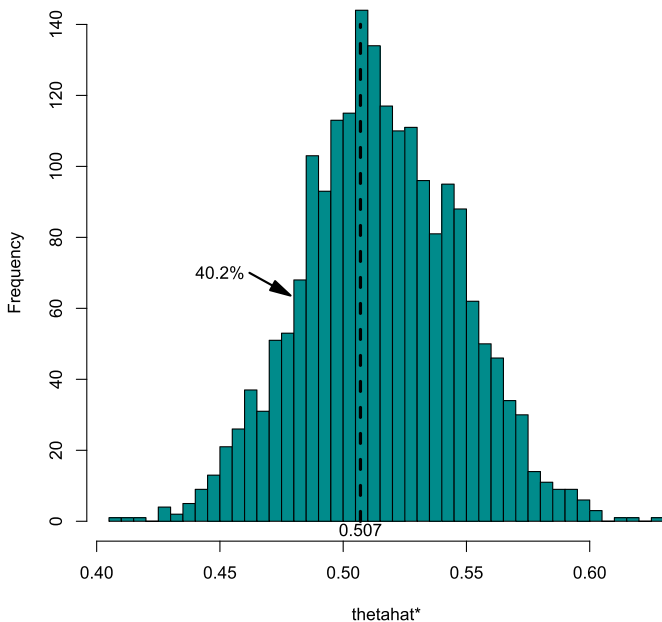
$$\hat{z}_0 = \Phi^{-1} \hat{G}(\hat{\theta}), \quad (2.9)$$

this following from  $z_0 = \Pr_{\phi}\{\hat{\phi} \leq \phi\}$  in (2.6) (assuming the true value  $\phi = 0$ ). Only 40.2% of the 2000  $\hat{\theta}^*$ 's were less than the original value  $\hat{\theta} = 0.507$ , indicating a substantial upward bias in  $\hat{\theta}$ . The estimated bias corrector

$$\hat{z}_0 = \Phi^{-1}(0.402) = -0.327 \quad (2.10)$$

**Table 1.** Lower and upper confidence limits for  $\theta$  having observed  $\hat{\theta} = 1$  from the model  $\hat{\theta} \sim \theta \cdot \text{Gamma}_{10}/10$ .

$\alpha$	Standard	pctile	bc	bca	Exact
0.025	0.38	0.48	0.52	0.585	0.585
0.16	0.69	0.69	0.74	0.764	0.764
0.84	1.31	1.31	1.39	1.448	1.448
0.975	1.62	1.71	1.80	2.086	2.085



**Figure 2.** 2000 nonparametric bootstrap replications of the adjusted  $R^2$  statistic for the diabetes example of Figure 1. Only 40.2% of the 2000  $\hat{\theta}^*$  values were less than  $\hat{\theta} = 0.507$ , suggesting its strong upward bias.

accounts for most of the downward shift of the bca limits seen in Figure 1.

The acceleration  $a$  cannot in general be read from the bootstrap distribution  $\hat{G}$ , this being an impediment to automatic bootstrap confidence intervals. The next two chapters include discussions of the computation of  $a$  in nonparametric and parametric situations, with further discussion in the Appendix.

### 3. Nonparametric bca Intervals

Nonparametric inference begins with an observed dataset

$$\mathbf{x} = (x_1, x_2, \dots, x_n), \quad (3.1)$$

where the  $x_i$  are independent and identically distributed (iid) observations from an unknown probability distribution  $F$  on a space  $\mathcal{X}$ ,

$$x_i \stackrel{\text{ind}}{\sim} F, \quad i = 1, 2, \dots, n. \quad (3.2)$$

The space  $\mathcal{X}$  can be anything at all—one-dimensional, multidimensional, functional, discrete—with no parametric form assumed for  $F$ . The assumption that  $F$  is the same for all  $i$  makes (3.2) a *one-sample problem*, the only type we will consider here.

A real-valued statistic  $\hat{\theta}$  has been computed by applying some estimating algorithm  $t(\cdot)$  to  $\mathbf{x}$ ,

$$\hat{\theta} = t(\mathbf{x}); \quad (3.3)$$

we wish to assign a confidence interval to  $\hat{\theta}$ . In the diabetes example of Section 1,  $x_i$  is  $v_i$  (1.2),  $n = 442$ ,  $\mathbf{x}$  is what was called  $\mathbf{v}$ , and  $\hat{\theta} = t(\mathbf{x})$  is the adjusted  $R^2$  statistic (1.3)–(1.4).

A nonparametric bootstrap sample  $\mathbf{x}^*$ ,

$$\mathbf{x}^* = (x_1^*, x_2^*, \dots, x_n^*), \quad (3.4)$$

is composed of  $n$  random draws, *with replacement*, from the set of original observations  $\{x_1, x_2, \dots, x_n\}$ . It produces a bootstrap replication

$$\hat{\theta}^* = t(\mathbf{x}^*) \quad (3.5)$$

of  $\hat{\theta}$ .  $B$  such replications are independently drawn,

$$\hat{\theta}_i^* = t(\mathbf{x}_i^*), \quad i = 1, 2, \dots, B, \quad (3.6)$$

as shown in Figure 2 for the diabetes example,  $B = 2000$ . The bootstrap estimate of standard error is the empirical standard deviation of the  $\hat{\theta}_i^*$  values,

$$\hat{\sigma}_{\text{boot}} = \left[ \sum_{i=1}^B (\hat{\theta}_i^* - \hat{\theta}^*)^2 / (B - 1) \right]^{1/2}, \quad (3.7)$$

(where  $\hat{\theta}^* = \sum \hat{\theta}_i^* / B$ ), which equals 0.032 in Figure 2.

The *jackknife* provides another estimate of standard error: let  $\mathbf{x}_{(i)}$  be the dataset of size  $n - 1$  having  $x_i$  removed from  $\mathbf{x}$ , and  $\hat{\theta}_{(i)}$  the corresponding estimate

$$\hat{\theta}_{(i)} = t(\mathbf{x}_{(i)}). \quad (3.8)$$

Then

$$\hat{\sigma}_{\text{jack}} = \left[ \frac{n-1}{n} \sum_{i=1}^n (\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)})^2 \right]^{1/2}, \quad \left( \hat{\theta}_{(\cdot)} = \sum_{i=1}^n \hat{\theta}_{(i)} / n \right), \quad (3.9)$$



which equals 0.033 in Figure 2;  $\hat{\sigma}_{\text{jack}}$  is not used in the bca algorithm, but jackknife methodology plays an important role in the program `bcajack`.

The vector of bootstrap replications, say

$$\mathbf{t}^* = (\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_B^*), \quad (3.10)$$

provides estimates  $\hat{G}$  (its cdf) and  $\hat{z}_0$  (2.9). For nonparametric problems, the jackknife differences

$$d_i = \hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)} \quad (3.11)$$

provide an estimate of the acceleration  $a$ ,

$$\hat{a} = \frac{1}{6} \sum_{i=1}^n d_i^3 / \left( \sum_{i=1}^n d_i^2 \right)^{3/2}. \quad (3.12)$$

The call `bcajack(x, B, func)`—where `func` represents  $t(\mathbf{x})$  (3.2)—computes  $\hat{G}$ ,  $\hat{z}_0$ , and  $\hat{a}$ , and the bca limits obtained from (2.2).

Our programs `bcajack2` and `bcapar` use a different estimate of  $a$  based directly on bootstrap replications (see the Appendix). This allows an estimate “jsd” of variability for  $\hat{a}$ ; see Table 3 in Section 4. The likelihood-based confidence interval methods mentioned previously make corrections to standard intervals that involve quantities like  $\hat{z}_0$  and  $\hat{a}$ , so jsd-type calculations may be relevant there, too.

Contemporary sample sizes  $n$  can be large, or even *very* large, making the calculation of  $n$  jackknife values impractically slow. Instead, the observations can be collected into  $m$  groups each of size  $g = n/m$ , with  $X_k$  the collection of  $x_i$ ’s in group  $k$ . The statistic  $\hat{\theta} = t(\mathbf{x})$  can just as well be thought of as a function of  $\mathbf{X} = (X_1, X_2, \dots, X_m)$ , say

$$\hat{\theta} = T(\mathbf{X}), \quad (3.13)$$

with  $\mathbf{X}$  an iid sample taking values in the product space  $\mathcal{X}^g$ . Now only  $m$ , instead of  $n$ , recomputations are needed to evaluate  $\hat{a}$  from (3.8)–(3.12). (Note: `bcajack` does not require separate specification of  $T(\cdot)$ .) For the diabetes example, Figures 1 and 2 were recomputed using  $m = 40$ , with only minor changes from  $m = 442$ .

There are two sources of error in the confidence limits  $\hat{\theta}_{\text{bca}}(\alpha)$ : *sampling error*, due to the random selection of  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  from  $F$ , and *internal error*, due to the Monte Carlo selection of  $B$  bootstrap samples  $\mathbf{x}^*(1), \mathbf{x}^*(2), \dots, \mathbf{x}^*(B)$ . The latter error determines how large  $B$  must be.

A different jackknife calculation provides estimates of internal error, based just on the original  $B$  bootstrap replications. The  $B$ -vector  $\mathbf{t}^*$  (3.10) is randomly partitioned into  $J$  groups of size  $B/J$  each, say  $J = 10$ . Each group is deleted in turn, and  $\hat{\theta}_{\text{bca}}(\alpha)$  recomputed. Finally, the jackknife estimate of standard error (3.9) is calculated from the  $J$  values of  $\hat{\theta}_{\text{bca}}(\alpha)$ . The red vertical bars in Figure 1 indicate  $\pm$  one such jackknife standard error.

Table 2 shows part of the `bcajack` output for the adjusted  $R^2$  statistic, diabetes data, as in Figure 2. The internal standard

**Table 2.** Partial output of nonparametric bootstrap confidence interval program `bcajack` for adjusted  $R^2$  statistic, diabetes data, Figures 1 and 2.

	\$lims				
	bcalims	jacksd	standard	pct	
0.025	0.437	0.004	0.444	0.006	
0.05	0.446	0.003	0.454	0.013	
0.1	0.457	0.002	0.466	0.032	
0.16	0.465	0.003	0.475	0.060	
0.5	0.498	0.001	0.507	0.293	
0.84	0.529	0.002	0.538	0.672	
0.9	0.540	0.002	0.547	0.768	
0.95	0.550	0.002	0.559	0.862	
0.975	0.560	0.002	0.569	0.918	

	\$stats					\$ustats	
	$\hat{\theta}$	$\hat{\theta}_{\text{boot}}$	$\hat{z}_0$	$\hat{a}$	$\hat{\sigma}_{\text{jack}}$	ustat	sdu
est	0.507	0.032	−0.327	−0.007	0.033	0.496	0.038
jsd	0.000	0.001	0.028	0.000	0.000		

error estimates for the bca limits  $\hat{\theta}_{\text{bca}}(\alpha)$  are labeled “jacksd.” The column “pct” gives the percentiles of the bootstrap histogram in Figure 2 corresponding to each  $\hat{\theta}_{\text{bca}}(\alpha)$ ; for example,  $\hat{\theta}_{\text{bca}}(0.975)$  corresponded to the 90.2 percentile of the 2000  $\hat{\theta}^*$  values, that is, the 1804th largest value. Having pct very near 0 or 1 for some  $\alpha$  suggests instability in  $\hat{\theta}_{\text{bca}}(\alpha)$ .

Also shown are the estimates  $\hat{\theta}$ ,  $\hat{\theta}_{\text{boot}}$ ,  $\hat{z}_0$ ,  $\hat{a}$ , and  $\hat{\sigma}_{\text{jack}}$ , as well as their internal standard error estimates “jsd.” ( $\hat{\theta}$ ,  $\hat{a}$ , and  $\hat{\sigma}_{\text{jack}}$  do not involve bootstrap calculations, and so have jsd zero.) Acceleration  $\hat{a}$  is nearly zero, while the bias corrector  $\hat{z}_0$  is large, with a substantial jsd.

The argument  $B$  in `bcajack` can be replaced by  $\mathbf{t}^*$ , the vector of  $B$  bootstrap replications (3.10). Calculating the  $\hat{\theta}^*$  values separately is sometimes advantageous:

- In cases where one wishes to compare `bcajack` output in different settings, for instance, for different choices of the grouping parameter  $m$ .
- If there is interest in each of the components of a  $p$ -dimensional parameter  $\Theta = T(\mathbf{x})$ , a  $B \times p$  matrix  $\mathbf{T}^*$  of bootstrap replications can be calculated from a single run of nonparametric samples, and `bcajack` executed separately for each column of  $\mathbf{T}^*$ .
- The “hybrid method,” discussed in Section 4, allows  $\mathbf{t}^*$  to be drawn from parametric probability models, and then analyzed with `bcajack`.
- For large datasets and complicated statistics  $\hat{\theta}$ , distributed computation of the  $\hat{\theta}^*$  values may be necessary.

This last point brings up a limitation of `bcajack`. Even if the vector of bootstrap replications  $\mathbf{t}^*$  is provided to `bcajack`, the program still needs to execute  $t(\cdot)$  for the jackknife calculations (3.8)–(3.12); an alternative algorithm `bcajack2` entirely avoids step (3.8) (as discussed in the Appendix). It is able to provide nonparametric bca intervals from just the replication vector  $\mathbf{t}^*$ . (It also provides a slightly better estimate of internal error: `bcajack` cheats on the calculations of jacksd by keeping  $a$  fixed instead of recalculating it fold to fold, while `bcajack2` let’s  $a$  vary.) `Bcajack2` is designed for situations where the

<sup>2</sup>See the Appendix for how `bcajack` proceeds if  $m$  does not exactly divide  $n$ .

function  $\hat{\theta} = \text{tfun}(\mathbf{x})$  is awkward to implement within the `bca` algorithm.

Both `bcajack` and `bcajack2` also return “ustat,” a bias-corrected version of  $\hat{\theta}$ ,

$$\text{ustat} = 2\hat{\theta} - \text{mean}(\hat{\theta}^*), \quad (3.14)$$

so  $\text{ustat} = \hat{\theta} - (\text{mean}(\hat{\theta}^*) - \hat{\theta})$ . Also provided is “sdu,” an estimate of sampling error for `ustat`, based on the theory in Efron (2014). For the adjusted  $R^2$  statistic of Table 2,  $\text{ustat} = 0.496$  with standard error 0.038, compared to  $\hat{\theta} = 0.507$  with standard error 0.038, a downward adjustment of about half a standard error.

#### 4. Bootstrap Confidence Intervals for Parametric Problems

The `bca` method was originally designed for parametric estimation problems. We assume that the observed data  $\mathbf{y}$  has come from a parametric density function

$$f_{\alpha}(\mathbf{y}), \quad (4.1)$$

where  $\alpha$  is the unknown parameter vector. We wish to set confidence intervals for a one-dimensional parameter of interest

$$\theta = s(\alpha). \quad (4.2)$$

An estimate  $\hat{\alpha} = A(\mathbf{y})$ , perhaps its MLE, provides

$$\hat{\theta} = s(\hat{\alpha}), \quad (4.3)$$

a point estimate of  $\theta$ . By sampling from  $f_{\hat{\alpha}}(\cdot)$  we obtain parametric bootstrap resamples  $\mathbf{y}^*$ ,

$$\mathbf{y}^* \sim f_{\hat{\alpha}}(\cdot). \quad (4.4)$$

Each  $\mathbf{y}^*$  provides an  $\hat{\alpha}^* = A(\mathbf{y}^*)$  and then a bootstrap replication of  $\hat{\theta}$ ,

$$\hat{\theta}^* = s(\hat{\alpha}^*). \quad (4.5)$$

Carrying out (4.4)–(4.5) some large number  $B$  of times yields a bootstrap data vector  $\mathbf{t}^* = (\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_B^*)$ . `Bca` confidence limits (2.2) can be estimated beginning from  $\mathbf{t}^*$ , but we will see that there is one complication not present in nonparametric settings.

The parametric bootstrap confidence interval program `bcapar` assumes that  $f_{\alpha}(\mathbf{y})$  (4.1) is a member of a  $p$ -parameter exponential family

$$f_{\alpha}(\mathbf{y}) = e^{\alpha' \hat{\beta} - \psi(\alpha)} f_0(\mathbf{y}), \quad (4.6)$$

having  $\alpha$  as the  $p$ -dimensional *natural parameter* and

$$\hat{\beta} = \hat{\beta}(\mathbf{y}) \quad (4.7)$$

as the  $p$ -dimensional sufficient statistic. Its expected value, the *expectation parameter*

$$\beta = E_{\alpha} \{ \hat{\beta}(\mathbf{y}) \} \quad (4.8)$$

is a one-to-one function of  $\alpha$ ;  $\hat{\beta}$  and its bootstrap replications  $\hat{\beta}^*$  play a central role in `bcapar`.

The parameter of interest  $\theta = s(\alpha)$  can also be expressed as a function of  $\beta$ , say  $\theta = \tau(\beta)$ , and then  $\hat{\theta}$  written as a function of  $\mathbf{y}$ ,

$$\hat{\theta} = \tau(\hat{\beta}(\mathbf{y})) \equiv t(\mathbf{y}). \quad (4.9)$$

In familiar applications, either  $s(\hat{\alpha})$  or  $\tau(\hat{\beta})$  may have simple expressions, or sometimes neither. A key factor of `bcapar` is its ability to work directly with  $\hat{\theta} = t(\mathbf{y})$  without requiring specification of  $s(\cdot)$  or  $\tau(\cdot)$ ; see the Appendix.

As an important example of exponential family applications, consider the logistic regression model

$$\boldsymbol{\eta} = \mathbf{M}\alpha, \quad (4.10)$$

where  $\mathbf{M}$  is a specified  $n \times p$  structure matrix, and  $\boldsymbol{\eta}$  is an  $n$ -vector of logit parameters  $\eta_i = \log \pi_i / (1 - \pi_i)$ . We observe

$$y_i = \begin{cases} 1 & \text{with probability } \pi_i \\ 0 & 1 - \pi_i \end{cases} \quad (4.11)$$

independently for  $i = 1, 2, \dots, n$ . This is a  $p$ -parameter exponential family having natural parameter  $\alpha$ , sufficient statistic

$$\hat{\beta} = \mathbf{M}'\mathbf{y}, \quad (4.12)$$

and expectation parameter  $\beta = \mathbf{M}'\boldsymbol{\pi}$ . In the *neonate* application below,  $\theta = s(\alpha)$  is the first coordinate of  $\alpha$ .

The vector of bootstrap replicates  $\mathbf{t}^* = (\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_B^*)$  provides an estimated bootstrap cdf  $\hat{G}$  for substitution into (2.2), the formula for  $\hat{\theta}_{\text{bca}}(\alpha)$ , and also the estimated bias-corrector  $\hat{z}_0 = \Phi^{-1}\hat{G}(\hat{\theta})$  as at (2.9). This leaves the acceleration  $a$ . In parametric problems  $a$  depends on the choice of sufficient statistic  $\hat{\beta} = \hat{\beta}(\mathbf{y})$  (4.7); `bcapar` takes as input  $t_0 = t(\mathbf{y}) = \hat{\theta}$ , the  $B$ -vector  $\mathbf{t}^*$  of bootstrap replicates, and also  $\mathbf{b}^*$ , the  $B \times p$  matrix of bootstrap sufficient vectors

$$\mathbf{b}^* = (\hat{\beta}_1^*, \hat{\beta}_2^*, \dots, \hat{\beta}_B^*)'. \quad (4.13)$$

In the logistic regression setting (4.10)–(4.12), the estimate  $\hat{\alpha}$  gives  $\hat{\boldsymbol{\eta}} = \mathbf{M}\hat{\alpha}$  and the vector of estimated probabilities

$$\hat{\pi}_i = 1/(1 + e^{-\hat{\eta}_i}), \quad i = 1, 2, \dots, n. \quad (4.14)$$

A parametric bootstrap sample  $\mathbf{y}^* = (y_1^*, y_2^*, \dots, y_n^*)$  is generated from

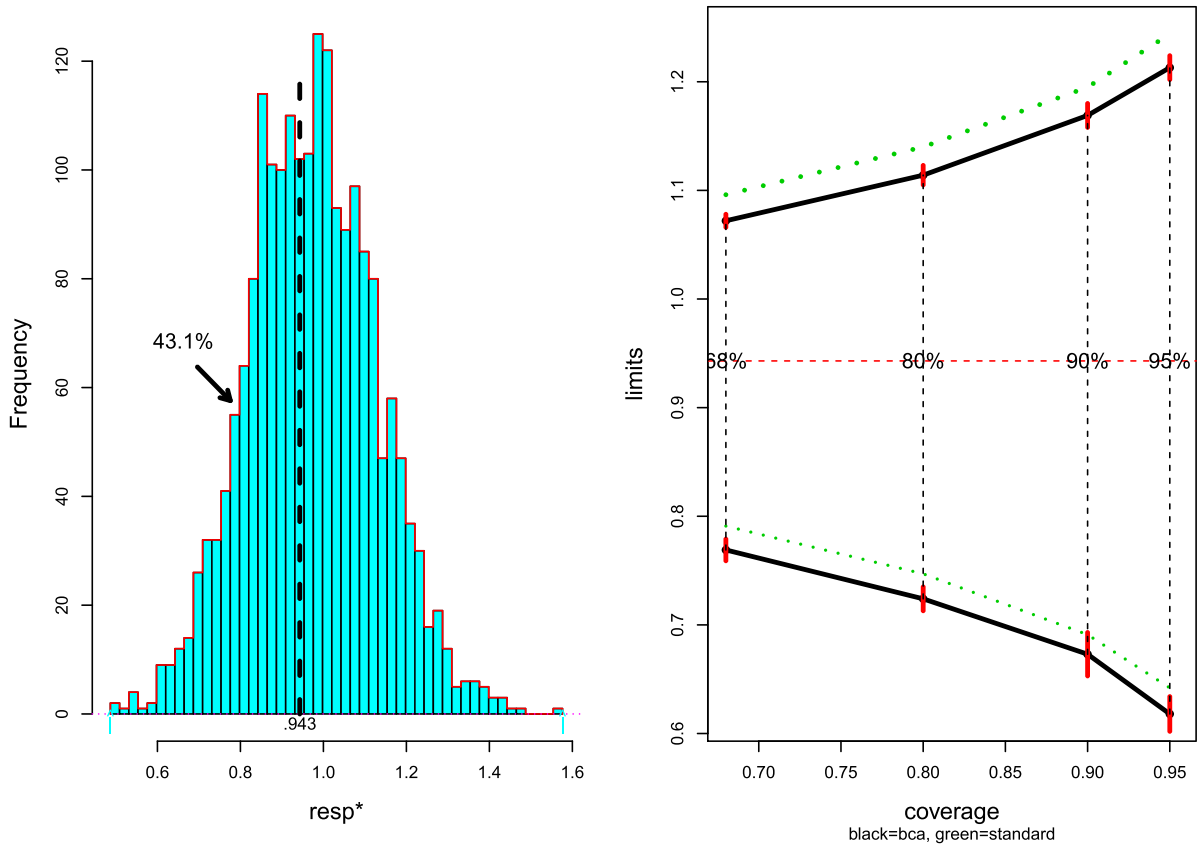
$$y_i^* = \begin{cases} 1 & \text{with probability } \hat{\pi}_i \\ 0 & 1 - \hat{\pi}_i \end{cases} \quad (4.15)$$

independently for  $i = 1, 2, \dots, n$ . Each  $\mathbf{y}^*$  provides a  $\hat{\theta}^* = t(\mathbf{y}^*)$  and also

$$\hat{\beta}^* = \mathbf{M}'\mathbf{y}^*, \quad (4.16)$$

the required inputs for `bcapar`.

Figure 3 refers to a logistic regression application: 812 neonates at a large clinic in a developing country were judged to be at serious risk; 205 died within 10 days, while 607 recovered. It was desired to predict death versus survival on the basis of 11 baseline variables, of which “resp,” a measure of respiratory distress, was of particular concern; resp was the first coordinate of  $\alpha$  so  $\hat{\theta} = s(\hat{\alpha}) = \hat{\alpha}_1$ .



**Figure 3.** Neonate data. *Left panel:* 2000 parametric bootstrap replications  $\widehat{\text{resp}}^*$ , MLE. *Right panel:* bca limits (solid) compared to standard limits (dotted), from `bccapar`. Estimates  $\hat{z}_0 = -0.215$ ,  $\hat{a} = -0.019$ .

A logistic regression analysis having  $M = 812 \times 11$  (with the columns of  $M$  standardized), yielded MLE  $t_0 = \hat{a}_1$  and bootstrap standard error for  $\text{resp}$ ,

$$\widehat{\text{resp}} = 0.943 \pm 0.155. \quad (4.17)$$

Program `bccapar`, using  $B = 2000$  bootstrap samples, provided the results pictured in Figure 3. The standard confidence limits for  $\text{resp}$  look reasonably accurate in this case, having only a small upward bias, about one-fifth of a standard error, compared to the bca limits, with almost all of the difference coming from the bias-corrector  $\hat{z}_0 = -0.215$  (with internal standard error 0.024).

Bootstrap methods are particularly useful for nonstandard estimation procedures. The neonate data were reanalyzed employing model (4.10)–(4.11) as before, but now estimating  $\alpha$  by the *glmnet* regression algorithm, a regularizing procedure that shrinks the components of  $\hat{\alpha}$ , some of them all the way to zero. The algorithm fits an increasing sequence of  $\alpha$  estimates, and selects a “best” one by cross-validation. Applied to the neonate data, it gave best  $\hat{\alpha}$  having

$$\widehat{\text{resp}} = 0.862 \pm 0.127. \quad (4.18)$$

Regularization reduced the standard error, compared with the MLE (4.15), but at the expense of a possible downward bias.

An application of `bccapar` made the downward bias evident. The left panel of Figure 4 shows that 66% of the  $B = 2000$  parametric bootstrap replicates  $\widehat{\text{resp}}^*$  were less than  $\widehat{\text{resp}} = 0.862$ , making the bias corrector large:  $\hat{z}_0 = 0.411$ . The right

panel plots the bca intervals, now shifted substantially upward from the standard intervals. Internal errors, the red bars, are considerable for the upper limits. (Increasing  $B$  to 4000 as a check gave almost the same limits.)

The two sets of bca confidence limits, logistic regression and *glmnet*, are compared in Figure 5. The centering of their two-sided intervals is about the same, *glmnet*’s large bias correction having compensated for its shrunken  $\widehat{\text{resp}}$  estimate (4.17), while its intervals are about 10% shorter.

As mentioned before, a convenient feature of `bccapar` that adds to the goal of automatic application is that both  $\hat{\theta}^*$  and  $\hat{\beta}^*$  can be directly evaluated as functions of  $y^*$ , without requiring a specific function from  $\hat{\beta}^*$  to  $\hat{\theta}^*$  (which would be awkward in a logistic regression setting, for example). If, however, a specific function is available, say

$$\hat{\theta}^* = \tau(\hat{\beta}^*), \quad (4.19)$$

then `bccapar` provides supplementary confidence limits based on the *abc* method (“approximate bootstrap confidence” intervals, as distinct from the more recent “approximate bayesian calculation” algorithm).

The *abc* method (DiCiccio and Efron 1992) substitutes local Taylor series approximations for bootstrap simulations in the calculation of bca-like intervals. It is very fast and, if  $s(\cdot)$  is smoothly defined, usually quite accurate. In a rough sense, it resembles taking  $B = \infty$  bootstrap replications. It does, however, require function (4.19) as well as other exponential



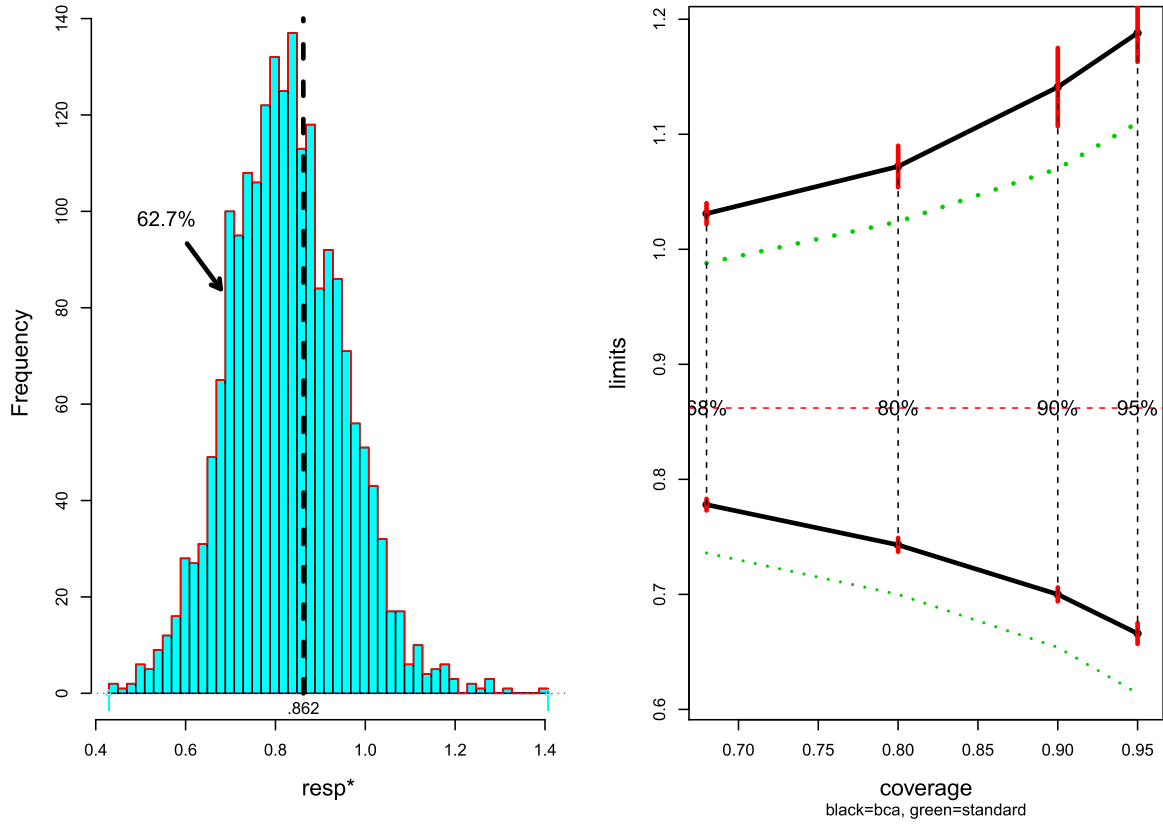


Figure 4. As in Figure 3, but now using glmnet estimation for  $\text{resp}$ , rather than logistic regression MLE.

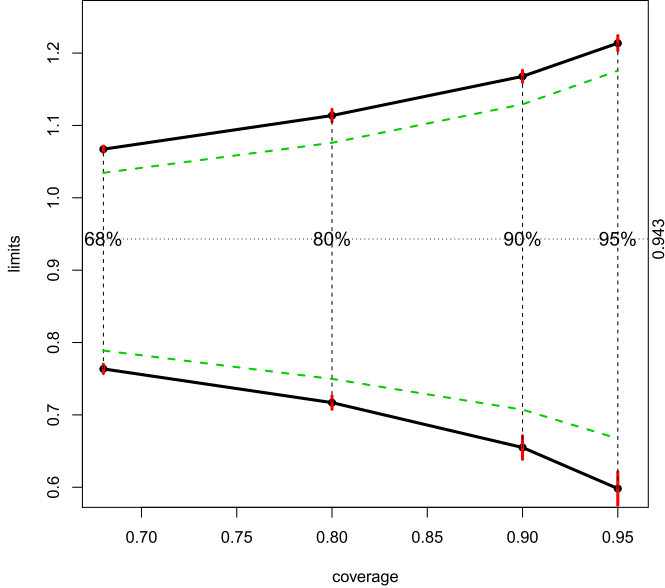


Figure 5. Comparison of bca confidence limits for neonate coefficient  $\text{resp}$ , logistic regression MLE (solid) and glmnet (dashed).

family specifications for each application, and is decidedly not automatic.

Program `abcpar`, discussed in the Appendix, uses the `bcapar` inputs  $t^*$  and  $b^*$  (4.13) to automate abc calculations. If function  $\tau(\cdot)$  is supplied to `bcapar`, it also returns the abc confidence limits as well as abc estimates of  $z_0$  and  $a$ . This is illustrated in the next example.

We suppose that two independent variance estimates have been observed in a normal theory setting,

$$\begin{aligned}\hat{\sigma}_1^2 &\sim \sigma_1^2 \chi_{n_1}^2/n_1, \\ \hat{\sigma}_2^2 &\sim \sigma_2^2 \chi_{n_2}^2/n_2,\end{aligned}\tag{4.20}$$

and that the parameter of interest is their ratio

$$\theta = \sigma_1^2/\sigma_2^2.\tag{4.21}$$

In this case the MLE  $\hat{\theta} = \hat{\sigma}_1^2/\hat{\sigma}_2^2$  has a scaled  $F$  distribution,

$$\hat{\theta} \sim \theta F_{n_1, n_2}.\tag{4.22}$$

Model (4.20) is a two-parameter exponential family having sufficient statistic

$$\hat{\beta} = (\hat{\sigma}_1^2, \hat{\sigma}_2^2).\tag{4.23}$$

Because (4.22) is a one-parameter scale family, we can compute *exact* level  $\alpha$  confidence limits for  $\theta$ ,

$$\hat{\theta}_{\text{exact}}(\alpha) = \hat{\theta}/F_{n_1, n_2}^{(1-\alpha)},\tag{4.24}$$

the notation indicating the  $1 - \alpha$  quantile of a  $F_{n_1, n_2}$  random variable.

`Bcapar` was applied to model (4.20)–(4.23), with

$$n_1 = 10 \quad \text{and} \quad n_2 = 42.\tag{4.25}$$

That is, parametric resamples were obtained according to

$$\begin{aligned}\hat{\sigma}_1^{2*} &\sim \hat{\sigma}_1^2 \chi_{n_1}^2/n_1, \\ \hat{\sigma}_2^{2*} &\sim \hat{\sigma}_2^2 \chi_{n_2}^2/n_2,\end{aligned}\tag{4.26}$$

**Table 3.** Bcapar output for ratio of normal theory variance estimates of  $\theta = \sigma_1^2/\sigma_2^2$  (4.20)–(4.21); bca limits are a close match to exact limits (4.24) given observed point estimate  $\hat{\theta} = 1$ .

	\$lims					
	bcalims	jacksd	pctiles	stand	abclims	exact.lims
0.025	0.420	0.007	0.074	−0.062	0.399	0.422
0.05	0.483	0.006	0.112	0.108	0.466	0.484
0.1	0.569	0.006	0.175	0.305	0.556	0.570
0.16	0.650	0.006	0.243	0.461	0.639	0.650
0.5	1.063	0.007	0.591	1.000	1.050	1.053
0.84	1.843	0.018	0.913	1.539	1.808	1.800
0.9	2.150	0.022	0.958	1.695	2.151	2.128
0.95	2.658	0.067	0.987	1.892	2.729	2.655
0.975	3.287	0.102	0.997	2.062	3.428	3.247

	\$stats					ustats			\$abcstats	
	$\theta$	$\hat{\sigma}_{boot}$	$a$	$z_0$	sd.delta	ustat	sdu	$B$	$a$	$z_0$
est	1	0.542	0.099	0.114	0.513	0.948	0.504	16000	0.1	0.1
jsd	0	0.004	0.004	0.010	0.004	0.004	0.005	0		

NOTE: Including `func=funcF` (4.28) in the call added abc limits and statistics.

**Table 4.** Actual coverage levels of the bca, abc, and standard limits shown in Table 3.

Nominal	bca	abc	Standard
0.025	0.024	0.021	0.110
0.05	0.050	0.046	0.140
0.1	0.097	0.097	0.187
0.16	0.150	0.158	0.237
0.5	0.493	0.503	0.541
0.84	0.840	0.849	0.960
0.9	0.901	0.909	0.997
0.95	0.951	0.958	1.000
0.975	0.976	0.982	NA

giving

$$\hat{\theta}^* = \hat{\sigma}_1^{2*}/\hat{\sigma}_2^{2*} \quad \text{and} \quad \hat{\beta}^* = (\hat{\sigma}_1^{2*}, \hat{\sigma}_2^{2*}). \quad (4.27)$$

The observed value  $\hat{\theta}$  simply scales the exact endpoints  $\hat{\theta}_{\text{exact}}(\alpha)$  and likewise  $\hat{\theta}_{\text{bca}}(\alpha)$ —correct behavior under transformations is a key property of the bca algorithm—so for comparison purposes the value of  $\hat{\theta}$  is irrelevant. It is taken to be  $\hat{\theta} = 1$  in what follows. Table 3 shows the output of bcapar based on  $B = 16,000$  replications of (4.27).  $B$  was chosen much larger than necessary in order to pinpoint the performance of bcapar vis-à-vis  $\hat{\theta}_{\text{exact}}(\alpha)$ . The match was excellent. Table 4 shows actual coverage levels for the endpoints of Table 3, bca almost exactly matching the nominal values.

Abc limits were also generated by including in the call, the function `ratio(.)`

$$\text{ratio}(\hat{\beta}^*) = \hat{\beta}_1^*/\hat{\beta}_2^*, \quad (4.28)$$

this being  $\tau(\hat{\beta}^*)$  in (4.19). The abc limits were less accurate for the more extreme values of  $\alpha$ . On the other hand, there are theoretical reasons for preferring the abc estimates of  $a$  and  $z_0$  in this case.

The standard, bca, and exact limits are graphed in the right panel of Figure 6. Two results stand out: bcapar has done an excellent job of matching the exact limits, and both methods suggest very large upward corrections to the standard limits. In this case, the three bca corrections all point in the same

direction:  $\hat{z}_0$  is positive,  $\hat{a}$  is positive, and  $\hat{G}$ , the bootstrap cdf, is long-tailed to the right, as seen in the left panel of Figure 6. (This last point relates to a question raised in the early bootstrap literature, as in Hall (1988): if  $\hat{G}$  is long-tailed to the *right* should not the confidence limits be skewed to the *left*? The answer is no, at least in exponential family contexts.)

Even with  $B = 16,000$  replications there is still a moderate amount of internal error in  $\hat{\theta}_{\text{bca}}(0.975)$ , as indicated by its red bar. The “pctiles” column of Table 3 suggests why:  $\hat{\theta}_{\text{bca}}(0.975)$  occurs at the 0.996 quantile of the  $B$  replications, that is, at the 64th largest  $\hat{\theta}^*$ , where there is a limited amount of data for estimating  $G(\cdot)$ . Bca endpoints are limited to the range of the observed  $\hat{\theta}^*$  values, and cannot be trusted when recipe (2.2) calls for extreme percentiles.

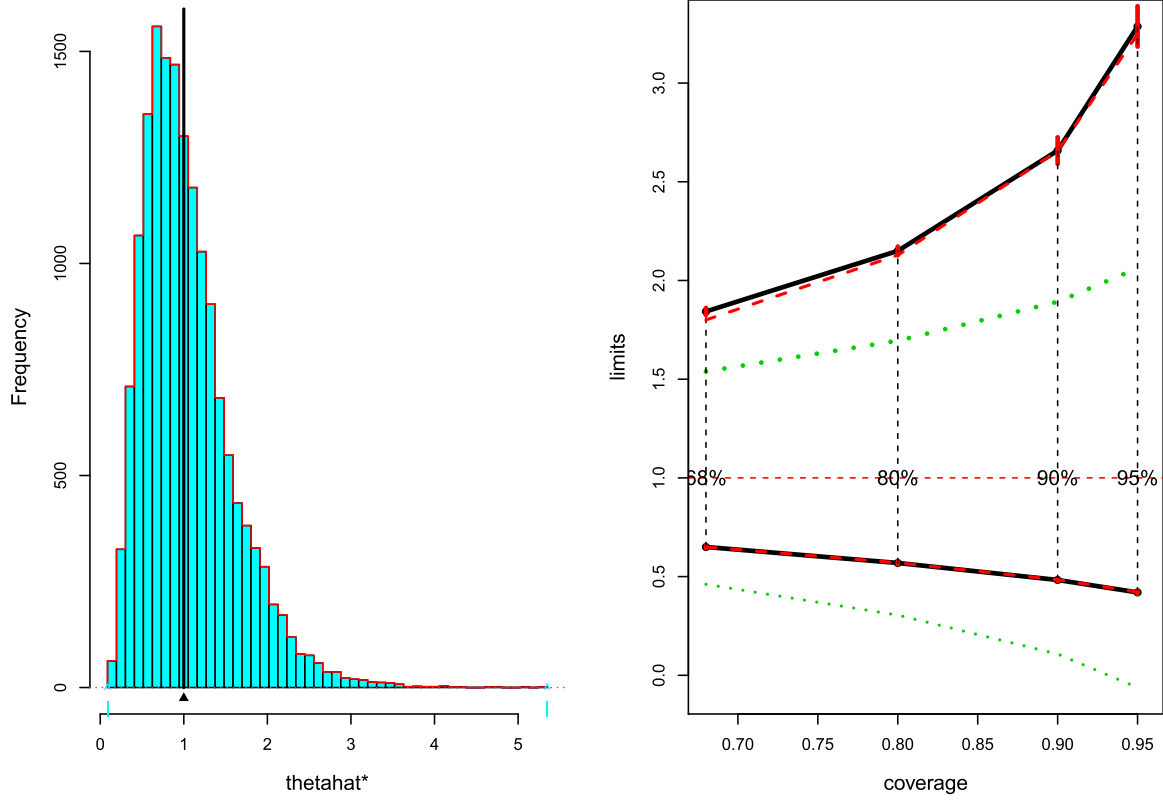
A necessary but “un-automatic” aspect of bcapar is the need to calculate the matrix of sufficient vectors  $\mathbf{b}^*$  (4.13). This can be avoided in one-sample situations by a *hybrid* application of bcapar–bcjack, at the risk of possible bias: the vector  $\mathbf{t}^*$  of  $B \hat{\theta}^*$  replications is drawn parametrically, as in the left panel of Figure 6. Then the call

$$\text{bcjack}(t_0 = \hat{\theta}, B = \mathbf{t}^*, \mathbf{r}) \quad (4.29)$$

uses  $\mathbf{t}^*$  to compute  $\hat{G}$  and  $\hat{z}_0$ , as in bcapar. In the neonate example,  $\mathbf{r}$  would take a 12-column matrix of predictors and responses, do a logistic regression or glmnet analysis of the last column on the first 11, and return the first coordinate of the estimated regression vector.

The bca analysis from (4.29) would differ only in the choice of acceleration  $a$  from a full use of bcapar. This made no difference in the neonate example,  $a$  being nearly 0 in both analyses.

If  $\theta$  is a component of the natural parameter vector of a multiparameter exponential family then the theoretically ideal confidence limits for  $\theta$  should be calculated conditionally on the other components of the MLE  $\hat{\alpha}$ . Bca limits are calculated unconditionally. DiCiccio and Young (2008) in parametric bootstrap procedures, concluded that



**Figure 6.** Left panel:  $B = 16,000$  bootstrap replications of  $\hat{\theta}^*$  (4.27). Right panel: confidence limits for  $\theta$  having observed  $\hat{\theta} = 1$ ; green standard, black bca, red dashed exact. The corrections to the standard limits are enormous in this case.

**Table 5.** Exact and approximate central 95% confidence intervals for  $\theta$  having observed  $x = 16$  from  $X \sim \text{Poisson}(\theta)$ .

	Lower	Upper
Exact	9.47	25.41
Pierce and Bellio	9.52	25.39
bcapar	9.82	25.53

the unconditional intervals can perform well, even from a conditional point of view. Pierce and Bellio (2017) aim at a tougher goal than us, computing second-order accurate confidence intervals that allow for conditioning. Their algorithm accomplishes this but at the expense of requiring more from the user, among other things profile likelihood computations, that is, for each possible value of the parameter of interest, the constructed MLE for the vector  $\partial P$  nuisance parameter.

As an easy example, when there are no nuisance parameters, we return to the case when we observe  $x = 16$  from a Poisson model with unknown mean  $\theta$ . The 95% central confidence limits are shown in Table 5.

## Appendix A

What follows are some additional comments on the new bca programs and the underlying theory.

### A.1. Bias Correction and Acceleration

The schematic diagram in Figure 7 relates to the estimates  $\hat{z}_0$  and  $\hat{a}$  in bcapar:  $\hat{\beta}$  is the MLE of the sufficient vector  $\beta$  in the exponential

family (4.6); it gives the estimate  $\hat{\theta} = \tau(\hat{\beta})$  (4.9) for the parameter of interest  $\theta$ .  $\hat{C}$  is the level surface of  $\beta$  vectors that give the same estimate of  $\theta$ ,

$$\hat{C} = \{\beta : \tau(\beta) = \hat{\theta}\}; \quad (\text{A.1})$$

$\dot{\tau}$  is the gradient vector of  $\tau(\beta)$  evaluated at  $\hat{\beta}$ ,

$$\dot{\tau} = (\partial \tau(\beta) / \partial \beta_j)_{\hat{\beta}}, \quad (\text{A.2})$$

and as such is orthogonal to  $\hat{C}$  at  $\hat{\beta}$ . The dots represent the bootstrap resamples  $\hat{\beta}_i^*$ ,  $i = 1, 2, \dots, B$ .

Let  $\hat{p}_z$  be the proportion of the  $\hat{\beta}_i^*$  lying below  $\hat{C}$  (i.e., in the direction opposite to  $\dot{\tau}$ ). Then assuming certain monotonicity properties,

$$\hat{z}_0 = \Phi^{-1}(\hat{p}_z), \quad (\text{A.3})$$

as in (2.9).

Define

$$D_i^* = (\hat{\beta}_i^* - \hat{\beta})^t \dot{\tau}, \quad (\text{A.4})$$

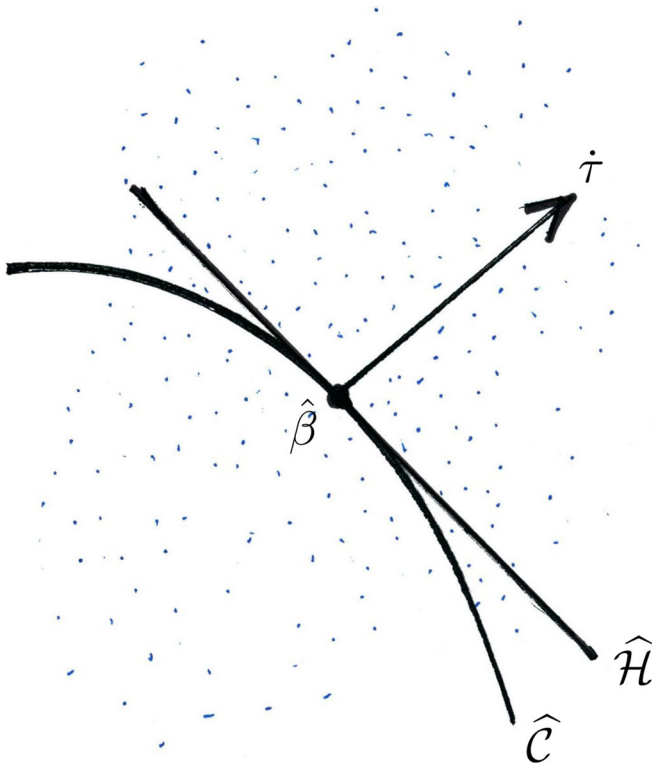
and let  $\hat{p}_a$  be the proportion of  $\hat{\beta}_i^*$  vectors lying below  $\hat{H}$ . A standard two-term Edgeworth expansion (Hall 1992, chap. 2) gives the approximation

$$\Phi^{-1}(\hat{p}_a) \doteq \hat{\gamma}/6, \quad (\text{A.5})$$

where  $\hat{\gamma}$  is the empirical skewness of the  $B$   $D_i^*$  values. It can also be shown, following (6.7) in Efron (1987), that the acceleration  $a$  equals  $\gamma/6$ , so that

$$\hat{a} \doteq \Phi^{-1}(\hat{p}_a) \doteq \hat{\gamma}/6. \quad (\text{A.6})$$

Both estimates of  $a$  are returned by bcapar,  $\hat{\gamma}/6$  labeled “a” and  $\Phi^{-1}(\hat{p}_a)$  labeled “az.” These were nearly the same in the examples of



**Figure 7.** Schematic diagram concerning estimation of the acceleration  $a$ , and its relation to the bias-corrector  $z_0$ , as explained in the text.

**Section 4.** Comparing (A.6) with (A.3), we can see that if  $\hat{C}$  curves away from  $\hat{\tau}$ , as in Figure 7, we would have  $\hat{z}_0 < \hat{a}$ . In a one-dimensional family, there is no curvature and  $\hat{z}_0 = \hat{a}$ .

How is the least favorable direction  $\hat{\tau}$  in Figure 7 calculated? This would be easy if the function  $\hat{\theta} = \tau(\hat{\beta})$  were available. However, *not* requiring the user to provide  $\tau(\hat{\beta})$  is essential to making `bcapar` easy to apply in practice. Only the original form of the statistic as a function of the observed data,  $\hat{\theta} = t(y)$ , is required. `Bcapar` uses local linear regression to estimate  $\hat{\tau}$ :

- The  $B \times p$  matrix  $\mathbf{b}^*$  (4.13) is standardized to, say,  $\mathbf{c}^*$ , having columns with mean 0 and variance 1.
- The rows  $c_i^*$  of  $\mathbf{c}^*$  are ranked in terms of their length.
- The index set  $\mathbf{I}^*$  of the lowest “Pct” quantile of lengths is identified; Pct = 0.333 by default.
- Finally,  $\hat{\tau}$  is set equal to the vector of ordinary linear regression coefficients of  $\hat{\theta}_i^*$  on  $c_i^*$ , within the set  $\mathbf{I}^*$ .

Theoretically, this estimate will approach  $\hat{\tau}$  as Pct  $\rightarrow$  0. Empirically the calculation is not very sensitive to the choice of Pct.

Highly biased situations, indicated by very large values of the bias corrector  $\hat{z}_0$ , destabilize the bca confidence limits. (Maximum likelihood estimation in high-dimensional models is susceptible to large biases.) The bias-corrected estimate `ustat` (3.14) can still be useful in such cases. `Bcapar`, like `bcajack`, returns `ustat` and its sampling error estimate “sdu,” and also estimates “jsd,” their internal errors. These errors are seen to be quite small in the example of Table 3.

`Bcapar` includes an option for also returning the bca confidence density, a form of posterior inference for  $\theta$  given the data based on ideas of both Fisher and Neyman; see Section 11.6 of Efron and Hastie (2016) and, for a broader picture, Xie and Singh (2013).

## A.2. `bcajack` and `bcajack2`

In the nonparametric setting,  $\hat{\tau}$  can be closely approximated in terms of the jackknife differences  $d_i = \hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}$  (3.11) by taking the  $i$ th component to be

$$\hat{\tau}_i = d_i \sqrt{n \cdot (n-1)} \quad (\text{A.7})$$

(Efron and Hastie 2016, formula (10.12)).

As discussed near the end of Section 3, using (A.7) in the calculation of the acceleration  $a$  makes `bcajack` require inputting the function for  $t(\mathbf{x})$  (3.8), even if the resamples  $\mathbf{t}^* = (\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_B^*)$  have been previously calculated. `Bcajack2` instead uses a regression estimate for  $\hat{\tau}$  and  $\hat{a}$ , analogous to that in `bcapar`.

Nonparametric resamples  $\hat{\theta}^* = t(\mathbf{x}^*)$  (3.5) can also be expressed as

$$\hat{\theta}^* = T(Y), \quad (\text{A.8})$$

$Y$  the count vector  $Y = (Y_1, Y_2, \dots, Y_n)$ , where, if  $\mathbf{x} = (x_1, x_2, \dots, x_n)$ ,

$$Y_i = \#\{x_j^* = x_i\}. \quad (\text{A.9})$$

The  $B \times n$  matrix  $\hat{Y}$  of count vectors for the  $B$  bootstrap samples plays the same role as  $\mathbf{b}^*$  (4.13) in parametric applications. `Bcajack2` regresses the  $\hat{\theta}_i^*$  values on  $Y_i$  for a subset of the  $B$  cases having  $Y_i$  near  $(1, 1, \dots, 1)$ , obtaining an estimate of  $\hat{\tau}$  as in `bcapar`. A previously calculated list of the vector  $\mathbf{t}^*$  and the corresponding matrix  $\hat{Y}$  can be entered into `bcajack2`, without requiring further specification of  $t(\mathbf{x})$ .

## A.3. `abcpar`

The original abc algorithm (DiCiccio and Efron 1992) required, in addition to the function  $\theta = \tau(\beta)$  (4.19), a function mapping the expectation parameter  $\beta$  (4.8) back to the natural parameter  $\alpha$  (4.6), say

$$\alpha = \mu(\beta). \quad (\text{A.10})$$

If  $\tau(\cdot)$  is provided to `bcapar`, as at (4.28), it calculates an approximate formula  $\alpha = \hat{\mu}(\beta)$  using the “empirical exponential family” (Efron 2014, sec. 6) and then returns abc estimates, as in Table 3.

The abc method is *local*, in the sense of only using resamples  $\hat{\beta}^*$  nearby to  $\hat{\beta}$ , which can be an advantage if  $\hat{\beta}$  falls into an unstable neighborhood of  $\hat{\theta} = \tau(\hat{\beta})$ . In any case, its results depend less on the resample size  $B$ , and provide a check on `bcapar`.

## Supplementary Materials

**Title:** `bcaboot`—Bias corrected bootstrap confidence intervals. Available on the Comprehensive R Archive Network (CRAN).

**Script to reproduce figures:** R scripts to reproduce figures in this article (`scripts.zip`).

## Funding

Research supported in part by National Science Foundation award DMS 1608182 (Bradley Efron). Research supported in part by the Clinical and Translational Science Award 1UL1 RR025744 for the Stanford Center for Clinical and Translational Education and Research (Spectrum) from the National Center for Research Resources, National Institutes of Health and award LM07033 (Balasubramanian Narasimhan).

## References

- Barndorff-Nielsen, O. (1983), "On a Formula for the Distribution of the Maximum Likelihood Estimator," *Biometrika*, 70, 343–365. [3]
- Barndorff-Nielsen, O. E., and Cox, D. R. (1994), *Inference and Asymptotics*, Monographs on Statistics and Applied Probability (Vol. 52), London: Chapman & Hall. [3]
- DiCiccio, T., and Efron, B. (1992), "More Accurate Confidence Intervals in Exponential Families," *Biometrika*, 79, 231–245. [3,7,11]
- (1996), "Bootstrap Confidence Intervals," *Statistical Science*, 11, 189–228. [1,3]
- DiCiccio, T. J., and Young, G. A. (2008), "Conditional Properties of Unconditional Parametric Bootstrap Procedures for Inference in Exponential Families," *Biometrika*, 95, 747–758. [3,9]
- Efron, B. (1987), "Better Bootstrap Confidence Intervals" (with comments and a rejoinder by the author), *Journal of the American Statistical Association*, 82, 171–200. [1,3,10]
- (2004), "The Estimation of Prediction Error: Covariance Penalties and Cross-Validation" (with comments and a rejoinder by the author), *Journal of the American Statistical Association*, 99, 619–642. [2]
- (2014), "Estimation and Accuracy After Model Selection," *Journal of the American Statistical Association*, 109, 991–1007. [6,11]
- Efron, B., and Hastie, T. (2016), *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*, Institute of Mathematical Statistics Monographs (Book 5), Cambridge: Cambridge University Press. [3,11]
- Hall, P. (1988), "Theoretical Comparison of Bootstrap Confidence Intervals" (with a discussion and a reply by the author), *The Annals of Statistics*, 16, 927–985. [3,9]
- (1992), *The Bootstrap and Edgeworth Expansion*, Springer Series in Statistics, New York: Springer-Verlag. [3,10]
- Pierce, D. A., and Bellio, R. (2017), "Modern Likelihood-Frequentist Inference," *International Statistical Review*, 85, 519–541. [3,10]
- Skovgaard, I. M. (1985), "Large Deviation Approximations for Maximum Likelihood Estimators," *Probability and Mathematical Statistics*, 6, 89–107. [3]
- Xie, M., and Singh, K. (2013), "Confidence Distribution, the Frequentist Distribution Estimator of a Parameter: A Review" (with discussion), *International Statistical Review*, 81, 3–39. [11]