SEVEN STAGES OF BOOTSTRAP

Rudolf Beran
Department of Statistics, University of California, Berkeley
Berkeley, CA 94720 U.S.A.

Key Words: coverage probability, iterated bootstrap, two-step bootstrap, Monte Carlo, high-dimensional parameter.

Abstract

This essay is organized around the theoretical and computational problem of constructing bootstrap confidence sets, with forays into related topics. The seven section headings are: Introduction; The Bootstrap World; Bootstrap Confidence Sets; Computing Bootstrap Confidence Sets; Quality of Bootstrap Confidence Sets; Iterated and Two-step Bootstrap; Further Resources.

1. Introduction

Bradley Efron's 1979 paper on the bootstrap in Statistics gained the immediate interest of his peers for several historical reasons. First, the bootstrap promised to extend formal statistical inference to situations too complex for existing methodology. By the late 1970's, theoretical statisticians had recognized that classical formulations of statistics, whether frequentist or Bayesian or otherwise, did not provide a reasonable way to analyze the large data sets arising in the computer age. This awkward defensive position made theoreticians receptive to the bootstrap, as well as to other data analytic ideas that seemed less model-dependent than classical statistical theory.

Second, by the 1970's, developments in theoretical statistics had provided tools that soon proved powerful in analysing the behavior of bootstrap procedures. For instance, the theory of robust statistics accustomed researchers to working with continuous or differentiable statistical functionals. This prepared the way for the later interpretation of bootstrap distributions as statistical functionals. The need to quantify contamination neighborhoods in robustness studies drew attention to metrics for probability measures. Huber (1981) presents both developments in robust statistics. Several decades of work on asymptotic optimality theory, culminating in the early 1970's with the local asymptotic minimax bound and with Hájek's convolution theorem, encouraged statisticians to think about weak convergence of triangular arrays. Ibragimov and Has'minskii (1981) give a comprehensive account. Edgeworth expansions and saddlepoint approximations saw a revival in the 1970's that is summarized by Hall (1992). These various theoretical ideas were well-suited to studying the convergence in probability of bootstrap distributions and the asymptotic properties of bootstrap procedures.

Third, bootstrap-like methods were natural as computers proliferated. From the 1960's onwards, some data analysts, not all statisticians, began experimenting with Monte Carlo simulations from fitted distributions. These resampling experiments were based more on intuition rather than on logical analysis and were published outside the main-stream statistical journals. Since a careful historical study has not yet been done, it is possible that the origins of the resampling idea are substantially older. (After all, the paired comparisons design in Statistics can be traced back to the philosopher Carneades, head of the Academy in Athens around 150 B. C., who argued that the different fortunes of twins disproves the efficacy of astrology). An essential contribution of Efron's (1979) paper was to formulate the bootstrap idea, as an intellectual object that could be studied theoretically, and give it a name.

The purpose of this essay is to introduce the bootstrap, to indicate when and in what sense it works, to discuss basic questions of implementation, and to illustrate the main points by example. The exposition is organized around the construction of bootstrap confidence sets—an application where bootstrap methods already enjoy considerable success—with forays into related topics. Section 7 contains suggestions for further reading.

2. The Bootstrap World

We recall that a statistical model for a sample $X = (X_1, \ldots, X_n)$ consists of a family of distributions, written $\{P_{n,\theta}: \theta \in \Theta\}$. One member of this model, the true distribution, is considered to generate probability samples similar to the observed data. However, the value of the parameter θ that identifies the true distribution is not known to the statistician. We suppose that the parameter space Θ is metric, but do not require it to be finite dimensional.

Bootstrap methods are a particular application of simulation ideas to the problem of statistical inference. From the sample X, we construct an estimator $\hat{\theta}_n = \hat{\theta}_n(X)$ that converges in $P_{n,\theta}$-probability to θ, for some convergence concept to be chosen. The bootstrap idea is then to:

- Create an artificial world in which the true parameter value is $\hat{\theta}_n$ and the sample X^* is generated from the fitted model $P_{n,\hat{\theta}_n}$. That is, the conditional distribution of X^*, given the data X, is $P_{n,\hat{\theta}_n}$.
- Act as if sampling distributions computed in the artificial world are accurate approximations to the corresponding true (but unknown) sampling distributions.

The *original world* of the statistician's model consists of the observable X whose distribution is $P_{n,\theta}$. The *bootstrap world* consists of the observable X^* whose conditional distribution, given X, is $P_{n,\hat{\theta}_n}$. In the original world, the distribution of X is unknown. However, in the bootstrap world, the distribution of X^* is fully known. Thus, any sampling distribution in the bootstrap world can be computed, at least in principle.

This brief description omits several important issues. First, for each statistical model, there may be many possible bootstrap worlds, each corresponding to a different choice of the estimator $\hat{\theta}_n$. Only some choices may be successful. Second, the plug-in method for constructing the model distribution in the bootstrap world can be generalized, and sometimes must be. When a high or infinite dimensional θ lacks a consistent estimator in a natural metric, it may still be possible to construct a useful bootstrap world that mimics only relevant aspects of the model in the original world. Time series analysis and curve estimation provide leading examples; see Mammen (1992) and Janas (1993) as well as Example 1 in Section 3. Third, computation of sampling distributions in the bootstrap world often involves Monte Carlo approximations, whose design raises further issues. Fourth, bootstrap methods are rarely exact; their theoretical justification typically rests on asymptotics under which the bootstrap world converges to the original world. These points will be developed further as the essay proceeds.

3. Bootstrap Confidence Sets

Suppose we wish to construct a confidence set for the parametric function $\tau = \tau(\theta)$. Classical theory advises us to find a pivot—a function of the sample X and of τ whose distribution under the model $P_{n,\theta}$ is continuous and completely known. Archetypal are confidence intervals for the mean of a $N(\mu, \sigma^2)$

distribution when location μ and scale σ are unknown. Here $\theta = (\mu, \sigma)$, the parametric function $\tau(\theta) = \mu$, and the pivot is the t-statistic, whose sampling distribution does not depend on the unknown θ. Though important as an ideal case, the exact pivotal technique is rarely available. It already fails to generate confidence intervals for the difference of two normal means in the Behrens-Fisher problem, for lack of a pivot.

Bootstrap ideas permit generalizing the pivotal method. Let $R_n(X, \tau)$ be a function of the sample and of τ, whose distribution under the model $P_{n,\theta}$ is denoted by $H_n(\theta)$. Because it need not be a pivot, but plays an analogous role, we call R_n a *root.* A plausible estimator of the root's sampling distribution is then the *bootstrap distribution* $\hat{H}_{n,B} = H_n(\hat{\theta}_n)$. This bootstrap distribution has two complementary mathematical interpretations:

- As defined, $\hat{H}_{n,B}$ is a random probability measure, the natural plug-in estimator of the sampling distribution of $R_n(X, \tau)$. From this viewpoint, $\hat{H}_{n,B}$ is a statistical functional that depends on the sample only through $\hat{\theta}_n$.
- Alternatively, $\hat{H}_{n,B}$ is the conditional distribution of $R_n(X^*, \tau(\hat{\theta}_n))$ given the sample X. In other words, $\hat{H}_{n,B}$ is the distribution of the root R_n in the bootstrap world described at the end of Section 2.

The interpretation as conditional distribution leads readily to Monte Carlo approximations for a bootstrap distribution (see Section 4). The interpretation as statistical functional is the starting point in developing asymptotic theory for bootstrap procedures, as we shall see next.

Suppose that, for some convergence concept in the parameter space Θ, both of the following conditions hold, for every $\theta \in \Theta$:

A. The estimator $\hat{\theta}_n$ converges in probability to θ as n increases.

B. For any sequence $\{\theta_n\}$ that converges to θ, the sampling distribution $H_n(\theta_n)$ converges weakly to the limit $H(\theta)$.

Then, the bootstrap distribution $\hat{H}_{n,B}$ also converges weakly, in probability, to the limit distribution $H(\theta)$. Though apparently very simple, this reasoning provides a template for checking the consistency of bootstrap estimators. The skill lies in choosing the convergence concept so as to achieve both conditions A and B.

We can now construct bootstrap confidence sets by analogy with the classical pivotal method. Let $\hat{H}_{n,B}^{-1}(\alpha)$ denote the α-th quantile of the bootstrap distribution and let T denote the space of possible values for the parametric function $\tau = \tau(\theta)$. Define the *bootstrap confidence set* for τ to be

$$C_{n,B} = \{t \in T: R_n(X,t) \leq \hat{H}_{n,B}^{-1}(\alpha)\}. \tag{3.1}$$

If conditions A and B above hold and if the limiting distribution $H(\theta)$ is continuous at its α-th quantile, then the coverage probability $P_{n,\theta}(C_{n,B} \ni \theta)$ converges to α as n tends to infinity. The following application to Stein confidence sets illustrates two key aspects of the bootstrap method: its remarkable power and the care often needed to harness this power when the dimension of θ is high relative to sample size.

EXAMPLE 1. We observe the time-series $X = (X_1, \ldots, X_n)$, which is related to the signal $\theta = (\theta_1, \ldots, \theta_n)$ by the following model: the distribution of X is normal with mean vector θ and with covariance matrix identity. The parametric function τ of interest is the signal θ itself. The classical confidence set of level α for θ is a sphere centered at X, with radius determined by the chi-squared distribution having n degrees of freedom. Let $|\cdot|$ denote Euclidean norm. A Stein confidence set is a sphere centered at the Stein estimator

$$\hat{\theta}_{n,S} = [1 - (n-2)/|X|^2]X. \tag{3.2}$$

The root that is used to determine the radius of a Stein confidence set is

$$R_n(X,\theta) = n^{-1/2}\{|\hat{\theta}_{n,S} - \theta|^2 - [n - (n-2)^2/|X|^2]\}, \tag{3.3}$$

which compares the loss of the Stein estimator with an unbiased estimator of its risk. This approach to confidence sets for θ was proposed at the end of Stein (1981). By invariance under the orthogonal group, the sampling distribution of the root (3.3) depends on θ only through $|\theta|$, and so may be written in the form $H_n(|\theta|^2/n)$.

Let $\{\theta_n \in R^n, n \geq 1\}$ denote any sequence such that $|\theta_n|^2/n \to a$, a finite non-negative constant. Then $H_n(|\theta_n|^2/n)$ converges weakly to a normal distribution with mean 0 and variance

$$\sigma^2(a) = 2 - 4a/(1+a)^2. \tag{3.4}$$

This is condition B for this example. To meet condition A requires a careful choice of the estimator of θ, such as

$$\hat{\theta}_{n,CL} = [1 - (n-2)/|X|^2]_+^{1/2} X. \tag{3.5}$$

Note the square root in (3.5), unlike in (3.2). The essential point is that, under the sequence $\{\theta_n\}$ described above, the estimators $\{|\hat{\theta}_{n,CL}|^2/n\}$ converge in probability to a, the limiting value of $\{|\theta_n|^2/n\}$. Consequently, the bootstrap distribution $\hat{H}_{n,B} = H_n(|\hat{\theta}_{n,CL}|^2/n)$ converges to the same $N(0, \sigma^2(a))$ limit as does the actual sampling distribution of the root.

On the other hand, the plausible alternative estimators $H_n(|\hat{\theta}_{n,S}|^2/n)$ and $H_n(|X|^2/n)$ both converge weakly, in probability, to the wrong limits (Beran, 1993). In the successful bootstrap world for this problem, the conditional distribution of X^* is $N(\hat{\theta}_{n,CL}, I)$, not $N(X, I)$ or $N(\hat{\theta}_{n,S}, I)$.

The bootstrap confidence set $C_{n,B}$ in this example is just the sphere centered at the Stein estimator $\hat{\theta}_{n,S}$ with radius

$$\hat{d}_{n,B} = [n - (n-2)^2/|X|^2 + n^{1/2}\hat{H}_{n,B}^{-1}(\alpha)]_+^{1/2}. \tag{3.6}$$

By the reasoning sketched above, the coverage probability of this bootstrap Stein confidence set is asymptotically α, in the uniform sense that

$$\lim_{n\to\infty} \sup_{|\theta|^2 \le nc} |P_{n,\theta}(C_{n,B} \ni \theta) - \alpha| = 0 \tag{3.7}$$

for every positive finite c. For more on bootstrap Stein confidence sets, see Beran (1993).

A very different approach to constructing bootstrap confidence sets is Efron's BC_a method. This is suited to one-dimensional parametric functions τ. The asymptotic relationship between the BC_a method and the root-based method described above is discussed in Hall (1992).

4. Computing Bootstrap Confidence Sets

Only rarely does a bootstrap distribution $\hat{H}_{n,B}$ have a closed form distribution. Strategies for computing the quantile $\hat{H}_{n,B}^{-1}$ fall into two broad categories: Monte Carlo approximations on the one hand; Edgeworth expansions or saddlepoint approximations on the other hand. Computers are potentially useful in doing the algebra of the analytic approximations as well as in performing Monte Carlo simulations. However, the computational emphasis to date has been on Monte Carlo algorithms.

The simplest, and very general, Monte Carlo approach is to construct, in the bootstrap world, M conditionally independent repetitions $X_1^*, \ldots X_M^*$ of the original experiment. The conditional distribution of each bootstrap sample X_j^*, given X, is $P_{n,\hat{\theta}_n}$. The empirical distribution of the values $\{R_n(X_j^*, \hat{\theta}_n): 1 \le$

$j \leq M\}$ then converges to the theoretical bootstrap distribution $\hat{H}_{n,B}$ as M increases. This approximation technique, whose origins lie in Monte Carlo tests, is responsible for the name *resampling* method that is sometimes used imprecisely as a synonym for bootstrap method. In reality, resampling is only one of the ways to approximate a bootstrap distribution.

How many bootstrap samples should we use when resampling? The answer to this question is twofold, as was pointed out by Hall (1986). On the one hand, to achieve accurate coverage probability, we should choose the number of bootstrap samples M so that $k/(M+1) = \alpha$ for some integer k; and then use the k-th order statistic of the values $\{R_n(X_j^*, \hat{\theta}_n)\}$ as the critical value for the numerical implementation of $C_{n,B}$. Then, the coverage probability of this Monte Carlo version of $C_{n,B}$, evaluated under the joint distribution of the sample X and of the artificial samples $\{X_j^*: 1 \leq j \leq M\}$, is α plus a term that goes to zero as n increases. That coverage probability can be accurate for large values of n, when M is small but chosen as above, is useful in debugging a simulation study of bootstrap confidence sets.

On the other hand, the Monte Carlo approximation to the theoretical confidence set $C_{n,B}$ is a randomized procedure. Unless M is large, the computed critical value, and consequently the computed confidence set, will depend strongly upon the realization of the artificial samples $\{X_j^*: 1 \leq j \leq M\}$. To limit the amount of randomization, writers on the bootstrap have moved, with time, from the suggestion that M be of order $O(10^2)$ to the recommendation that M be as large as possible and preferably at least of order $O(10^3)$.

Several authors have investigated more efficient Monte Carlo schemes for approximating bootstrap distributions. Most successful in the bootstrap context have been importance sampling (Johns, 1988), balanced resampling (Davison, Hinkley, and Schechtman, 1986), and antithetic sampling (Snijders, 1984). Appendix II of Hall (1992) compares the relative efficiencies, when M is large, of these methods for approximating a bootstrap distribution function or quantile.

The discussion above pretends that random number generators produce realizations of independent, identically distributed random variables. This assumption is, at best, a rough approximation. A more satisfactory analysis of Monte Carlo approximations to bootstrap confidence sets is an open problem.

Edgeworth approximations to bootstrap distributions have proved valuable in studying the asymptotic properties of bootstrap confidence sets (Hall, 1992). As a practical means for determining bootstrap critical values, Edgeworth expansions suffer from relative inaccuracy in their tails as well as algebraic cumbersomeness. Saddlepoint approximations to bootstrap distributions, initiated by Davison and Hinkley (1988), appear to be more accurate, but currently lack convenient implementation outside the simplest cases.

5. Quality of Bootstrap Confidence Sets

A good confidence set is both reliable and selective. By reliability, we mean that the coverage probability is accurate; by selectivity we mean that the confidence set is not too large. Keeping a confidence set small, among all those of coverage probability α, is a fundamental design question, a matter of picking the root well. Achieving accurate coverage probability is then the simpler matter of constructing a good critical value for the chosen root. General criteria for picking a root include: minimizing $P_{n,\theta}(C_{n,B} \ni \theta')$ for $\theta' \neq \theta$, as Neyman proposed; or minimizing a geometrical risk such as $E_\theta \sup\{|t-\theta|: t \in C_{n,B}\}$. The bootstrap Stein confidence set in Example 1 has smaller geometrical risk, at every α and for sufficiently large n, than does the classical confidence sphere centered at X (Beran, 1993).

Bootstrap theory has made significant progress in understanding how to control coverage probability once the root is chosen. A number of important examples exhibit the following structure: The left continuous distribution function $H_n(\cdot, \theta)$ of the root admits an asymptotic expansion

$$H_n(x,\theta) = H_A(x,\theta) + n^{-k/2}h(x,\theta) + O(n^{-(k+1)/2}), \tag{5.1}$$

where the first two terms on the right hand side are smooth functions of θ, k is a positive integer, and the asymptotic distribution function $H_A(x,\theta)$ is continuous and strictly monotone in x. In this setting, a competitor to the bootstrap confidence set $C_{n,B}$ is the *asymptotic confidence set* for τ:

$$C_{n,A} = \{t \in T: R_n(X,t) \le H_A^{-1}(\alpha, \hat{\theta}_n)\}. \tag{5.2}$$

Like $C_{n,B}$, the asymptotic coverage probability of $C_{n,A}$ is α.

To compare rates-of-convergence of the coverage probabilities to α, suppose that the estimators $\{\hat{\theta}_n\}$ are $n^{-1/2}$-consistent. By heuristic argument, as in Beran (1988b), we find:

- If the asymptotic distribution H_A of the root depends on θ, then the coverage probabilities of $C_{n,A}$ and $C_{n,B}$ converge to α at the same rates.

- If the asymptotic distribution H_A does *not* depend on θ, then the coverage probability of $C_{n,B}$ converges to α faster than does the coverage probability of $C_{n,A}$.

In the first case, both the asymptotic and bootstrap approaches estimate the leading term of the expansion (5.1). In the second case, the bootstrap approach successfully estimates the second term in the expansion (the leading term is now known); however the simple asymptotic approach continues to estimate only the first term, having no information about the second term. The asymptotic approach might be refined by using a two term Cornish-Fisher expansion to generate the critical value in (5.2). In practice, this refinement may not be easy. The bootstrap approach is attractively intelligent in its handling of both cases without technical intervention by the statistician. Hall (1992) has placed the heuristics above on a rigorous footing, in a certain more specialized setting.

EXAMPLE 2. As an instance of the case most favorable to bootstrapping, let us consider the Behrens-Fisher problem—devising a confidence interval for the difference between two means when the variances in two independent normal samples are unknown and possibly unequal. We take as root the t-statistic constructed from the difference of the two sample means. The limiting distribution of this root, under the normal model, is standard normal. Bootstrapping from the fitted normal model for the two samples yields a confidence set that is asymptotically equivalent and numerically close to Welch's solution (Beran, 1988b). Moreover, if n denotes the combined sample size, the error in coverage probability of both the Welch and the bootstrap confidence sets is of order $O(n^{-2})$. By contrast, the asymptotic confidence set based on the normal limiting distribution of the t-statistic incurs a coverage probability error of order $O(n^{-1})$.

EXAMPLE 1 (continued). In this Stein confidence set problem, the limiting normal distribution of the root depends upon the unknown parameter through the limiting value of $|\theta|^2/n$. The asymptotic variance of the root (3.3) is estimated consistently by

$$\hat{\sigma}_n^2 = \sigma^2(|\hat{\theta}_{n,CL}|^2/n) \tag{5.3}$$

for σ^2 defined in (3.4). The bootstrap Stein confidence set $C_{n,B}$ was described in Section 3. The corresponding asymptotic Stein confidence set is the sphere centered at $\hat{\theta}_{n,S}$ with radius

$$\hat{d}_{n,A} = [n - (n-2)^2/|X|^2 + n^{1/2}\hat{\sigma}_n\Phi^{-1}(\alpha)]_+^{1/2}. \tag{5.4}$$

Here the coverage probability errors of $C_{n,A}$ and $C_{n,B}$ are both of order $O(n^{-1/2})$, as shown in Beran (1993). Figure 1 plots, for $n = 19$, the coverage probabilities of $C_{n,A}$ (diamonds) and $C_{n,B}$ (crosses) against the normalized noncentrality parameter $|\theta|^2/n$. The intended coverage probability is $\alpha = .90$; each bootstrap critical value is computed from 199 bootstrap samples by the method described in Section 4; and the coverage probabilities themselves are estimates based on 20,000 pseudo-random normal samples. The marked changes that occur in coverage probability as the normalized noncentrality parameter increases from 0 to 2 reflect variations in the asymptotic skewness and in the slope of the asymptotic variance of the root.

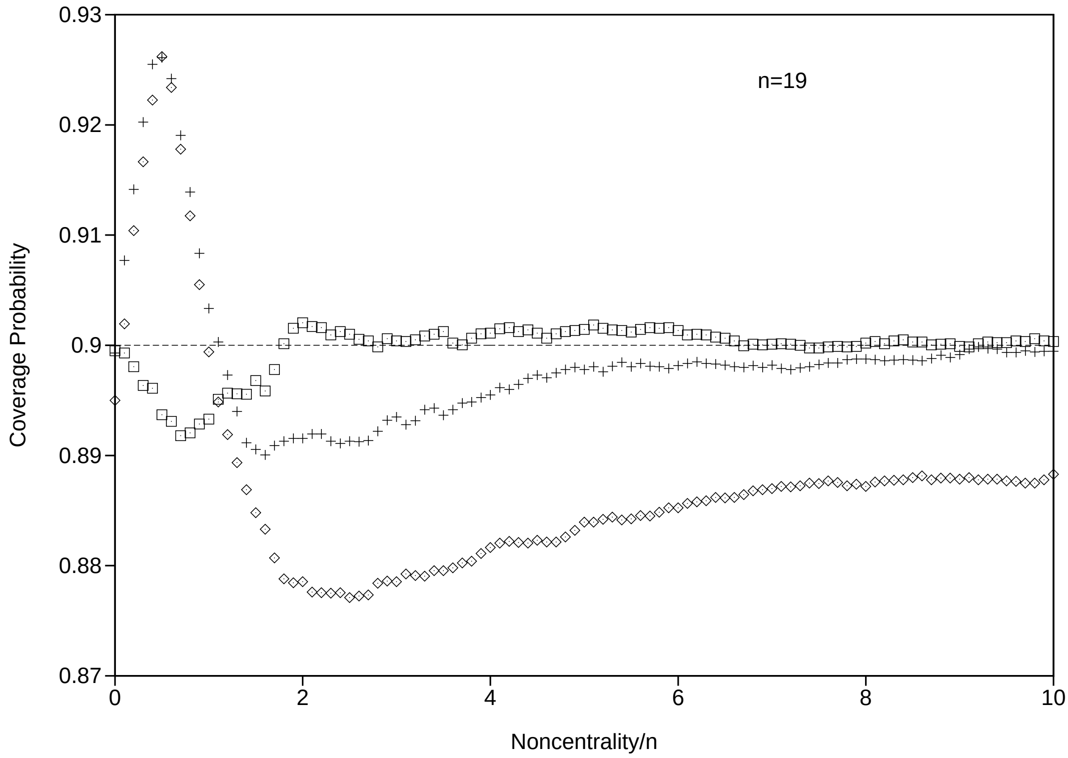

FIGURE 1. Coverage probabilities in Example 1 of $C_{n,A}$ (diamonds), of $C_{n,B}$ (crosses), and of $C_{n,TB}$ (squares) when α is .90 and n is 19.

To improve coverage probability accuracy of the Stein confidence set $C_{n,B}$, we can pursue a more sophisticated strategy: First transform the root in a one-to-one way so that its asymptotic distribution does *not* depend on the unknown parameter; and then construct the bootstrap confidence set based on the transformed root. Studentizing, as was done implicitly in Example 2, is an instance of such transformation. However, studentizing does not work well for moderate values of n in Example 1 or in other cases where the distribution of the root is substantially non-normal. More successful in Example 1 is the use of a variance stabilizing transformation. Instead of (3.3), consider the root

$$R_{n,T}(X,\theta) = n^{1/2}\{g[|\hat{\theta}_{n,S} - \theta|^2/n] - g[1-(n-2)^2/(n|X|^2)]\}, \tag{5.5}$$

where

$$g(u) = 2^{-1}\log[-2+4u+2^{3/2}(2u^2-2u+1)^{1/2}]. \tag{5.6}$$

The limiting distribution of root (5.5) is standard normal, in view of (3.4). Let $C_{n,TB}$ denote the transformed bootstrap Stein confidence set that is based on $R_{n,T}(X,\theta)$. The coverage probability error in $C_{n,TB}$ is of order $O(n^{-1})$, a significant improvement over $C_{n,A}$ and $C_{n,B}$ that is borne out by the coverage probabilities (squares) plotted in Figure 1.

6. Iterated and Two-step Bootstrap

We can use the bootstrap itself to transform a root $R_n(X,\tau)$ into a new root whose limiting distribution does not depend on the unknown parameter. Let $\hat{H}_{n,B}(\cdot)$ denote the left continuous bootstrap distribution function of the root R_n and define

$$R_{n,B}(X,\tau) = \hat{H}_{n,B}(R_n(X,\tau)) = H_n(R_n(X,\tau),\hat{\theta}_n). \tag{6.1}$$

When the limiting distribution of R_n is continuous, the limiting distribution of the new root $R_{n,B}$ is typically Uniform (0,1). Let $C_{n,BB}$ denote the bootstrap confidence set based on $R_{n,B}$. If $\hat{H}_{n,BB}$ denotes the bootstrap distribution of $R_{n,B}(X,\tau)$, then

$$C_{n,BB} = \{t \in T: R_n(X,t) \le \hat{H}_{n,B}^{-1}[\hat{H}_{n,BB}^{-1}(\alpha)]\}. \tag{6.2}$$

In the light of Section 5, we expect that the coverage probability of $C_{n,BB}$ converges to α at a faster rate than the coverage probability of $C_{n,B}$. This often turns out to be the case, as argued in Beran (1988b) and elsewhere. The transformation (6.1) is called *prepivoting*, because it maps the original root into one that is more nearly pivotal when n is large.

Construction of $C_{n,BB}$ involves two bootstrap worlds. In the *first bootstrap world*, as described in Section 2, the true parameter is $\hat{\theta}_n$ and we observe an artificial sample X^* whose conditional distribution, given X, is $P_{n,\hat{\theta}_n}$. Write θ_n^* for $\hat{\theta}_n(X^*)$, the recalculation of the estimator in the first bootstrap world. In the *second bootstrap world*, the true parameter is θ_n^* and we observe an artificial sample X^{**} whose conditional distribution, given X and X^*, is P_{n,θ_n^*}. Then

- The conditional distribution of $R_n^* = R_n(X^*, \tau(\hat{\theta}_n))$, given X, is the bootstrap distribution $\hat{H}_{n,B}$.
- The conditional distribution of $R_{n,B}^* = R_{n,B}(X^*, \tau(\hat{\theta}_n))$, given X, is the bootstrap distribution $\hat{H}_{n,BB}$. Moreover, by (6.1),

$$R_{n,B}^* = H_n(R_n^*, \theta_n^*) = P(R_n^{**} < R_n^* | X, X^*), \tag{6.3}$$

where $R_n^{**} = R_n(X^{**}, \tau(\theta_n^*))$.

From this we see that practical computation of $C_{n,BB}$ generally requires a double nested Monte Carlo algorithm. The inner level of this algorithm approximates $H_{n,B}$, while both levels are needed to approximate $H_{n,BB}$. For further details, see Beran (1988b). Constructing the second bootstrap world is often called iterated or double bootstrapping. The underlying idea is that differences between the first bootstrap world and the original world (which are unknown) approximately equal corresponding differences between the second bootstrap world and the first bootstrap world (which are computable).

Prepivoting is not the only use for iterated bootstrapping. Other inferential problems, such as bias reduction, can benefit from repeated bootstrapping, as discussed by Hall and Martin (1988). Alternative constructions of iterated bootstrap confidence sets, asymptotically equivalent to those derived from prepivoting, are treated by Hall (1992).

Superficially similar to double bootstrapping, but different logically and much less intensive computationally, is two-step bootstrapping. Two-step bootstrapping provides a way to extend the classical Tukey and Scheffé simultaneous confidence sets from normal linear models to general models. Suppose that the parametric function τ has *components* labelled by an index set U; that is $\tau(\theta) = \{\tau_u(\theta) : u \in U\}$. For each u, let $C_{n,u}$ denote a confidence set for the component τ_u. By simultaneously asserting the confidence sets $\{C_{n,u}\}$, we obtain a simultaneous confidence set C_n for the family of parametric functions $\{\tau_u\}$. The problem is to construct the component confidence sets $\{C_{n,u}\}$ in such a way that

$$P_{n,\theta}(C_{n,u} \ni \tau_u) \text{ is the same for every } u \in U \tag{6.4}$$

and

$$P_{n,\theta}(C_n \ni \tau) = \alpha. \tag{6.5}$$

Suppose that $R_{n,u} = R_{n,u}(X, \tau_u)$ is a root for the component parametric function τ_u. Let $H_{n,u}(\cdot, \theta)$ and $H_n(\cdot, \theta)$ denote the left-continuous distribution functions of $R_{n,u}$ and of $\sup_u H_{n,u}(R_{n,u}, \theta)$ respectively. The corresponding bootstrap estimators for these two distributions are then $\hat{H}_{n,u,B} = H_{n,u}(\cdot, \hat{\theta}_n)$ and $\hat{H}_{n,B} = H_n(\cdot, \hat{\theta}_n)$. Define the critical values

$$\hat{d}_{n,u} = \hat{H}_{n,u,B}^{-1}[\hat{H}_{n,B}^{-1}(\alpha)]. \tag{6.6}$$

Let T_u and T denote, respectively, the ranges of $\tau_u(\theta)$ and $\tau(\theta)$. Every point in the range set T can be written in component form $t = \{t_u\}$, where t_u lies in T_u. Define a bootstrap confidence set for τ_u by

$$C_{n,u,B} = \{t_u \in T_u \colon R_{n,u}(X, t_u) \le \hat{d}_{n,u}\}. \tag{6.7}$$

Simultaneously asserting these component confidence sets generates the following bootstrap simultaneous confidence set for τ:

$$C_{n,B} = \{t \in T \colon R_{n,u}(X, t_u) \le \hat{d}_{n,u} \text{ for every } u \in U\}. \tag{6.8}$$

Asymptotically in n, the confidence set $C_{n,B}$ satisfies the overall coverage probability condition (6.5); and the confidence sets $\{C_{n,u,B}\}$ satisfy the *balance* condition (6.4). Regularity conditions that ensure the validity of these conclusions are analogous to conditions A and B in Section 3. Beran (1988a) gives particulars. Interestingly, the Tukey and Scheffé simultaneous confidence intervals in the normal linear model are special cases of the bootstrap confidence set (6.8). These classical procedures satisfy (6.4) and (6.5) exactly.

Since the definition of simultaneous confidence set $C_{n,B}$ involves only the first bootstrap world, a Monte Carlo approximation to the critical values (6.7) requires only one round of resampling. Indeed, $\hat{H}_{n,u,B}$ and $\hat{H}_{n,B}$ are just the conditional distributions of $R_{n,u}(X^*, \tau_u(\hat{\theta}_n))$ and of $\sup_u H_{n,u}(R_{n,u}(X^*, \tau_u(\hat{\theta}_n), \hat{\theta}_n))$, given X. Computational difficulties can arise when the index set U is not finite. However, in practice we are usually interested in only a finite number of parametric functions. Iterated bootstrapping can be used to improve the rate at which the simultaneous confidence set approaches properties (6.4) and (6.5) as n increases. For details, see Beran (1990).

7. Further Resources

In this short account, we have sketched only how bootstrap methods may be used to construct reliable confidence sets. Significant progress has occurred in several additional directions, including: bootstrap tests; bootstrap prediction regions; bootstrap confidence sets for models where the dimension of the parameter space is high relative to sample size (Example 1 illustrates this situation); bootstrap inference based on nonparametric regression estimators or density estimators; bootstrap inference for spectral density estimators. Further information on these and other bootstrap developments may be found in the following sources:

Monographs. Efron and Tibshirani (1993) give a wide-ranging, relatively nonmathematical introduction to the bootstrap and its applications. Hall (1992) uses Edgeworth expansions to study higher-order asymptotic properties of bootstrap methods; the appendices treat other important aspects of bootstrap theory. Each chapter ends with brief bibliographical notes citing related work by other authors. Mammen (1992) develops higher-order bootstrap analyses without Edgeworth expansions; bootstrap worlds for models where the dimension of the parameter space is large relative to sample size (the *wild* bootstrap); and bootstrap methods for M-estimators in such circumstances. The dissertation of Janas (1993) covers bootstrap procedures based on the periodogram. Beran and Ducharme (1991) records six introductory lectures on bootstrap inference. Efron (1982) raises several problems that remain incompletely solved.

Survey papers. Surveys of bootstrap theory, which reflect the state of knowledge at the time of writing, include: Hinkley (1988), DiCiccio and Romano (1988), and Beran (1984). The Trier proceedings volume (Jöckel, Rothe, and Sendler, 1992) contains papers on random number generation as well as on bootstrap theory and applications. A second bootstrap proceedings volume is Billard and LePage (1992).

Acknowledgements

Grant support from the National Science Foundation, notably DMS 9224868, and the hospitality of Sonderforschungsbereich 123 at Universität Heidelberg gave me time to think about bootstrap methods over an extended period. This essay could not have been written otherwise.

References

Beran, R. (1984). Bootstrap methods in statistics. *Jber. d. Dt. Math. Verein.* **86**, 212–225.

Beran, R. (1988a). Balanced simultaneous confidence sets. *J. Amer. Statist. Assoc.* **83**, 679–686.

Beran, R. (1988b). Prepivoting test statistics: A bootstrap view of asymptotic refinements. *J. Amer. Statist. Assoc.* **83**, 686–697.

Beran, R. (1990). Refining bootstrap confidence sets. *J. Amer. Statist. Assoc.* **85**, 417–426.

Beran, R. (1993). Stein confidence sets and the bootstrap. Preprint.

Beran, R. and Ducharme, G. (1991). *Asymptotic Theory for Bootstrap Methods in Statistics.* Publications CRM, Université de Montréal.

Billard, L. and LePage. R. (eds.) (1992). *Exploring the Limits of Bootstrap.* Wiley, New York.

Davison, A. C. and Hinkley, D. V. (1988). Saddlepoint approximations in resampling methods. *Biometrika* **75**, 417–431.

Davison, A. C., Hinkley, D. V., and Schechtman, E. (1986). Efficient bootstrap simulation. *Biometrika* **73**, 555–566.

DiCiccio, T. J. and Romano, J. P. (1988). A review of bootstrap confidence intervals (with discussion). *J. Roy. Statist. Soc. Ser. B* **50**, 338–354.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Math. Statist.* **7**, 1–26.

Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans.* SIAM, Philadelphia.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Chapman and Hall, New York.

Hall, P. (1986). On the number of bootstrap simulations required to construct a confidence interval. *Ann. Statist.* **14**, 1453–1462.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer, New York.

Hall, P. and Martin, M. A. (1988). On bootstrap resampling and iteration. *Biometrika* **75**, 661–671.

Hinkley, D. V. (1988). Bootstrap methods (with discussion). *J. Roy. Statist. Soc. Ser. B* **50**, 321–337.

Huber, P. J. (1981). *Robust Statistics.* Wiley, New York.

Ibragimov, I. A. and Has'minskii, R. Z. (1981). *Statistical Estimation: Asymptotic Theory.* Springer, New York.

Janas, D. (1993). *Bootstrap Procedures for Time Series.* Shaker, Aachen.

Jöckel, K.-H., Rothe, G., and Sendler, W. (eds.) (1992). *Bootstrapping and Related Techniques.* Lecture Notes in Economics and Mathematical Systems **367**. Springer, Berlin.

Johns, M. V. Jr. (1988). Importance sampling for bootstrap confidence intervals. *J. Amer. Statist. Assoc.* **83**, 709–714.

Mammen, E. (1992). *When Does Bootstrap Work?* Lecture Notes in Statistics **77**. Springer, New York.

Snijders, T. A. B. (1984). Antithetic variates for Monte Carlo estimation of probabilities. *Statist. Neerland.* **38**, 55–73.

Stein, C. (1981). Estimation of the mean of a normal distribution. *Ann. Statist.* **9**, 1135-1151.