METRON - *International Journal of Statistics*
2003, vol. LXI, n. 2, pp. 227-242

G. ALASTAIR YOUNG

Better bootstrapping by constrained prepivoting

Summary - Bootstrap methods are attractive empirical procedures for assessment of errors in problems of statistical estimation, and allow highly accurate inference in a vast range of problems. Conventional bootstrapping involves sampling from the empirical distribution function in nonparametric problems, or a fitted parametric model in parametric inference. Recently, much attention has been focussed on methods for reduction of the error properties of bootstrap procedures, by systematic modification of the sampling model, in a way that is dependent on the parameter of interest. In this paper, we provide a general perspective on the bootstrap, based on the notion of prepivoting, with the specific aim of synthesizing recent developments related to modified, or "weighted", bootstrap procedures, and provide a critical evaluation of the practical benefits of such procedures over conventional bootstrap schemes and alternative analytic methods.

Key Words - Bootstrap; Conditional inference; Confidence set; Nonparametric bootstrap; Parametric bootstrap; Prepivoting; Tilted distribution; Weighted bootstrap.

1. INTRODUCTION

Since its introduction by Efron (1979), the bootstrap has become a method of choice for empirical assessment of errors and related quantities in a vast range of problems of statistical estimation. Bootstrap methodology encompasses a whole body of ideas, principal among them: (1) the *substitution principle*, of replacement in frequentist inference of an unknown probability distribution F by an estimate $\widetilde{F}$ constructed from the sample data; and (2) replacement of analytic calculation by *simulation* from $\widetilde{F}$. In conventional bootstrapping, $\widetilde{F}$ has a simple form. In nonparametric inference, $\widetilde{F}$ is the empirical distribution function $\widehat{F}$ of an observed random sample $Y = \{Y_1, \ldots, Y_n\}$, while in a parametric context a parametric model $F(y; \psi)$ with a parameter ψ of fixed dimension is replaced by its maximum likelihood estimate $F(y; \widehat{\psi})$. Much recent development in bootstrap methodology relates to the issue of whether systematic reductions in the error properties of bootstrap procedures may be obtained by simple modification of the sampling distribution $\widetilde{F}$.

Received April 2003 and revised June 2003.

In this paper, we provide a general perspective on the bootstrap paradigm, appropriate to both the parametric and nonparametric contexts, and describe how, in principle, the sought after systematic improvements may be obtained by constrained, or weighted, bootstrapping. Section 2 describes a general framework for an inference problem concerning an unknown scalar parameter of interest, and presents illustration of how common parametric and nonparametric inference procedures may be described within that framework. In Section 3 we revisit a view of the bootstrap due to Beran (1987), (1988), known as "prepivoting", and detail the operation of conventional bootstrapping from that perspective. Section 4 describes, from this prepivoting perspective, how better bootstrap procedures might be obtained, for both nonparametric and parametric problems, by constrained bootstrapping, and discusses various issues, practical and theoretical, related to such weighted bootstrapping. Numerical illustrations are presented in Section 5. Concluding remarks are made in Section 6, where we attempt a critical evaluation of the benefits of weighted bootstrap schemes over the conventional bootstrap, and provide thoughts for future developments.

2. An inference problem

Suppose that $Y = \{Y_1, \ldots, Y_n\}$ is a random sample from an unknown underlying distribution F, and let $\gamma \equiv \gamma(F)$ be a scalar parameter of interest.

Let $u(Y, \gamma)$ be a function of the data sample Y and the unknown parameter γ, such that a one-sided confidence set of nominal coverage $1 - \alpha$ for γ is $\mathcal{I} = \{\psi : u(Y, \psi) \leq 1 - \alpha\}$. We speak of $u(Y, \gamma)$ as a "confidence set root". A notational point is of importance here. In our development, we will denote by γ the *true* parameter value, with ψ denoting a generic point in the parameter space, a "candidate value" for inclusion in the confidence set.

We now provide two examples, the first parametric and the second relating to nonparametric inference about γ.

Example 1: Signed root likelihood ratio statistic. Suppose that it may be assumed that Y has probability density $f_Y(y; \theta)$ belonging to a specified parametric family, depending on an unknown parameter vector $\theta = (\gamma, \xi)$, with nuisance parameter ξ. Inference about γ may be based on the profile log-likelihood $l_p(\gamma) = l(\gamma, \widehat{\xi}_\gamma)$, and the associated likelihood ratio statistic $w_p(\gamma) = 2\{l_p(\widehat{\gamma}) - l_p(\gamma)\}$, with $l(\gamma, \xi) = \log f_Y(y; \gamma, \xi)$ the log-likelihood, $\widehat{\theta} = (\widehat{\gamma}, \widehat{\xi})$ the overall maximum likelihood estimator of θ, and $\widehat{\xi}_\gamma$ the constrained maximum likelihood estimator of ξ, for fixed γ.

As the parameter of interest is scalar, inference is conveniently based on the signed root likelihood ratio statistic, $r_p(\gamma) = \mathrm{sgn}(\widehat{\gamma} - \gamma) w_p(\gamma)^{1/2}$. We have that r_p is distributed as $N(0, 1)$ to error of order $O(n^{-1/2})$, and therefore a

confidence set of nominal coverage $1-\alpha$ for γ is $\{\psi : u(Y, \psi) \leq 1-\alpha\}$, with

$$u(Y, \psi) = \Phi\{r_p(\psi)\}\,.$$

Monotonicity in ψ of $u(Y, \psi)$ implies that the confidence set is of the form $(\hat{\gamma}_l, \infty)$, where the lower confidence limit $\hat{\gamma}_l$ is obtained by solving $\Phi\{r_p(\psi)\} = 1–\alpha$. The coverage error of the confidence set is of order $O(n^{-1/2})$. The error may be reduced to order $O(n^{-3/2})$ by analytically adjusted versions of r_p of the form

$$r_a = r_p + r_p^{-1} \log(u_p / r_p)\,,$$

that are distributed as $N(0, 1)$ to error of order $O(n^{-3/2})$: see, for example, Barndorff-Nielsen (1986). Here the statistic u_p depends on specification of an ancillary statistic, a function of the minimal sufficient statistic that is approximately distribution constant.

Example 2: Studentized parameter estimate. Let $\hat{\gamma}$ be an asymptotically normal nonparametric estimator of γ, with estimated variance $\hat{\sigma}^2$. A nonparametric confidence set of nominal coverage $1-\alpha$ may be defined as above by

$$u(Y, \psi) = \Phi\{(\hat{\gamma} - \psi)/\hat{\sigma}\}\,.$$

Again, the confidence set is of one-sided form $(\hat{\gamma}_l, \infty)$, with $u(Y, \hat{\gamma}_l) = 1-\alpha$. Typically, coverage error is again $O(n^{-1/2})$.

3. The prepivoting view of the bootstrap

From the prepivoting perspective (Beran, 1987, 1988), the bootstrap may be viewed as a device by which we attempt to transform the confidence set root $U = u(Y, \gamma)$ into an approximate pivot, that is, an approximately $\mathrm{Un}(0, 1)$ random variable.

The underlying notion is that if U were exactly distributed as $\mathrm{Un}(0, 1)$, the confidence set would have coverage exactly equal to $1-\alpha$: $\Pr(\gamma \in \mathcal{I}) = \Pr\{u(Y, \gamma) \leq 1-\alpha\} = \Pr\{\mathrm{Un}(0, 1) \leq 1-\alpha\} = 1-\alpha$. But U is typically *not* $\mathrm{Un}(0, 1)$, so the coverage error of $\mathcal{I}$ is non-zero. By bootstrapping, we hope to produce a new confidence set root u_1 so that the confidence set $\{\psi : u_1(Y, \psi) \leq 1-\alpha\}$ has lower coverage error for γ. The error properties of different bootstrap schemes can be assessed by measuring how close to uniformity is the distribution of $U_1 = u_1(Y, \gamma)$.

In the conventional bootstrap approach, the distribution function $G(x; \psi)$ of $u(Y, \psi)$ is estimated by

$$\hat{G}(x) = \Pr^*\{u(Y^*, \hat{\gamma}) \leq x\}\,,$$

and we define the conventional prepivoted root by

$$\hat{u}_1(Y, \psi) = \hat{G}\{u(Y, \psi)\}\,,$$

for each candidate parameter value ψ.

In a parametric problem, $\Pr^*$ denotes the probability under the drawing of bootstrap samples Y^* from the fitted maximum likelihood model $f_Y(y;\widehat{\theta})$. In the nonparametric setting, $\Pr^*$ denotes the probability under the drawing of bootstrap samples Y^* from the empirical distribution function $\widehat{F}$: such a sample is obtained by independently sampling, with replacement, from $\{Y_1, \ldots, Y_n\}$. In practice, in both contexts, the prepivoting must in general be carried out by performing a Monte Carlo simulation, involving the drawing of a series of R bootstrap samples, rather than analytically. In this regard, we note that the effect of Monte Carlo approximation on the coverage properties of the confidence set can be subtle (Lee and Young, 1999a), though a rule-of-thumb would suggest taking R to be of the order of a few thousands.

The basic idea here is that if the bootstrap estimated the sampling distribution exactly, so that $\widehat{G}$ was the true (continuous) distribution function G of $u(Y,\gamma)$, then $\widehat{u}_1(Y,\gamma)$ would be exactly $\mathrm{Un}(0,1)$ in distribution, as a consequence of the probability integral transform: if Z is a random variable with continuous distribution function $H(\cdot)$, then $H(Z)$ is distributed as $\mathrm{Un}(0,1)$. Therefore the confidence set $\{\psi : \widehat{u}_1(Y,\psi) \leq 1-\alpha\}$ would have exactly the desired coverage. Use of $\widehat{G}$ in place of G incurs an error, though in general the error associated with $\widehat{u}_1(Y,\psi)$ is smaller in magnitude than that obtained from $u(Y,\psi)$.

In Example 1, conventional bootstrapping amounts to replacing the asymptotic $N(0,1)$ distribution of r_p by its distribution when the true parameter value is $\widehat{\theta} = (\widehat{\gamma}, \widehat{\xi})$. The bootstrap confidence set is of the form $(\widehat{\gamma}_l^*, \infty)$, where $r_p(\widehat{\gamma}_l^*) = \widehat{c}_{1-\alpha}$, with $\widehat{c}_{1-\alpha}$ denoting the $1-\alpha$ quantile of $r_p(\widehat{\gamma})$ under sampling from the specified model with parameter value $(\widehat{\gamma}, \widehat{\xi})$. In general, this reduces the order of the coverage error of the confidence set to $O(n^{-1})$. That the conventional bootstrap approximates the true distribution of r_p to error of order $O(n^{-1})$ was established by DiCiccio and Romano (1995). The effectiveness of bootstrapping in approximation of the sampling distribution of the likelihood ratio statistic w_p was shown by Martin (1990), while Bickel and Ghosh (1990) demonstrated that bootstrap approximation automatically yields Bartlett correction of w_p.

In Example 2, prepivoting by the same means amounts to replacing the asymptotic $N(0,1)$ distribution of $(\widehat{\gamma}-\gamma)/\widehat{\sigma}$ by the distribution of $(\widehat{\gamma}^*-\widehat{\gamma})/\widehat{\sigma}^*$, with $\widehat{\gamma}^*$ and $\widehat{\sigma}^*$ denoting the estimator and its standard error estimator respectively for a bootstrap sample obtained by uniform resampling from $\{Y_1, \ldots, Y_n\}$. The confidence set is again of the form $(\widehat{\gamma}_l^*, \infty)$, where $\widehat{u}_1(Y, \widehat{\gamma}_l^*) = 1-\alpha$. In general, the bootstrapping again reduces the order of the coverage error to $O(n^{-1})$.

4. BETTER BOOTSTRAPS

The basic device by which we may, in principle, obtain better bootstrap inference is to change the distribution from which bootstrap samples are drawn. Specifically, instead of using a single distribution, we constrain or weight the distribution from which bootstrap samples are drawn, prescribing it to depend on the candidate value ψ of the parameter of interest.

Weighted bootstrap procedures of this kind encompass a variety of statistical methods and are closely related in the nonparametric setting to empirical and other forms of nonparametric likelihood (Owen, 1988, DiCiccio and Romano, 1990). Besides the inference problem of confidence set construction (and the associated hypothesis testing problem) considered here, applications of weighted bootstrap ideas are numerous, and include variance stabilization, nonparametric curve estimation, nonparametric sensitivity analysis etc.: see Hall and Presnell (1999a,b,c,).

In detail, in our prepivoting formulation of bootstrapping, we replace $\hat{u}_1(Y, \psi)$ by the weighted prepivoted root

$$\tilde{u}_1(Y, \psi) = \widetilde{G}\{u(Y, \psi); \psi\},$$

with

$$\widetilde{G}(x; \psi) = \Pr^\dagger\{u(Y^\dagger, \psi) \le x\}.$$

Now, in the parametric setting, $\Pr^\dagger$ denotes the probability under the drawing of bootstrap samples $Y^\dagger$ from the constrained fitted model $f_Y(y; \psi, \widehat{\xi}_\psi)$. In a nonparametric problem, $\Pr^\dagger$ denotes the probability under the drawing of bootstrap samples $Y^\dagger$ from the distribution $\widehat{F}_p$ which places probability mass p_i on Y_i, where $p \equiv p(\psi) = (p_1, \ldots, p_n)$ is chosen to minimize (say) the Kullback-Liebler distance

$$-n^{-1} \sum_{i=1}^{n} \log(np_i)$$

between $\widehat{F}_p$ and $\widehat{F}$, subject to $\gamma(\widehat{F}_p) = \psi$.

A number of remarks are in order.

Remark 1. We note that, by contrast with the conventional bootstrap approach, in principle at least, a different fitted distribution is required for each candidate parameter value ψ. In the context of Example 1, for instance, the confidence set is $\{\psi : r_p(\psi) \le c_{1-\alpha}(\psi, \widehat{\xi}_\psi)\}$, where now $c_{1-\alpha}(\psi, \widehat{\xi}_\psi)$ denotes the $1 - \alpha$ quantile of the sampling distribution of $r_p(\psi)$ when the true parameter value is $(\psi, \widehat{\xi}_\psi)$, so that a different bootstrap quantile is applied for each candidate ψ.

Remark 2. However, computational shortcuts which reduce the demands of weighted bootstrapping are possible. These include the use of stochastic search

procedures, which allow construction of the confidence set without a costly simulation at each candidate parameter value, such as the Robbins-Munro procedure (Garthwaite and Buckland, 1992; Carpenter, 1999), and, in the nonparametric case, approximation to the probability weights $p(\psi)$ (Davison and Hinkley, 1997, Section 9.4.1), rather than explicit evaluation for each ψ.

Remark 3. The theoretical effects of weighted bootstrapping in the nonparametric context are analyzed for various classes of problem, including those involving robust estimators and regression estimation, as well as the smooth function model of Hall (1992), by Lee and Young (2003). The basic conclusion is striking: if $u(Y, \gamma)$ is uniform to order $O(n^{-j/2})$,

$$\Pr\{u(Y, \gamma) \leq u\} = u + O(n^{-j/2}),$$

then, quite generally, $\widehat{u}_1(Y, \gamma)$ is uniform to order $O(n^{-(j+1)/2})$, while $\tilde{u}_1(Y, \gamma)$ is uniform to the higher order $O(n^{-(j+2)/2})$. The result holds for confidence set roots of the kind described in Example 2, as well as more complicated roots: an example is the widely used bootstrap percentile method confidence set root $u(Y, \gamma) = \Pr^*(\widehat{\gamma}^* > \gamma)$.

The basic assumption made by Lee and Young (2003) is that the root $u(Y, \gamma)$ admits an asymptotic expansion of the form

$$u(Y, \gamma) = \Phi(T) + \phi(T)\{n^{-1/2} r_1(\overline{Z}, T) + n^{-1} r_2(\overline{Z}, T) + \ldots\}, \qquad (1)$$

where $T = (\widehat{\gamma} - \gamma)/\widehat{\sigma}$ is the studentized parameter estimate, asymptotically standard normal, as in Example 2 above, and where the precise specification of $\overline{Z} = n^{-1} \sum_{i=1}^{n} z_i(Y, F)$ and polynomials r_1, r_2 depends on the class of problem being considered. The basic result then holds under mild conditions on the choice of probability weights p_i. In particular, the conclusions hold for a whole class of distance measures which generalize the Kullback-Leibler distance (Baggerly, 1998; Corcoran, 1998). The choice of distance measure is therefore largely irrelevant to the theoretical conclusion, allowing the use of well-developed algorithms (Owen, 2001) for construction of weighted bootstrap distributions $\widehat{F}_p$, as well as use of simple tilted forms of the empirical distribution function $\widehat{F}$, as described, for example, by DiCiccio and Romano (1990).

Lee and Young (2003) also consider the effects of successively iterating the prepivoting. They demonstrate that iterated weighted prepivoting accelerates the rate of convergence to zero of the bootstrap error, compared to the effect of iteration of the conventional bootstrap (Hall and Martin, 1988; Martin, 1990).

The same conclusions hold for testing. When testing a point null hypothesis $H_0 : \gamma = \gamma_0$, a one-sided test of nominal size α rejects H_0 if $u(Y, \gamma_0) \leq \alpha$. If $u(Y, \gamma_0)$ were exactly Un(0, 1), the null rejection probability would be exactly α. To increase accuracy, weighted bootstrapping applied with $\gamma = \gamma_0$ reduces

error by $O(n^{-1})$. Now, of course, weighted bootstrapping need only be carried out at the single value γ_0, so computational complications over conventional bootstrapping are reduced.

As an extension to previous work, we note here that the machinery developed by Lee and Young (2003) applies immediately to the signed root of empirical and other nonparametric likelihoods in the context of $M-$estimation. Considering the empirical likelihood case, DiCiccio and Monti (2001) provide an expansion, their formula (13), for the signed root $R(\gamma)$ of the empirical likelihood ratio statistic. It follows from this expansion that $\Phi\{R(\gamma)\}$ has an expansion of the form (1) above. Therefore, under the mild assumptions on the probability weights required by Lee and Young (2003), weighted prepivoting of $\Phi\{R(\gamma)\}$ reduces the coverage error of confidence sets for γ from $O(n^{-1/2})$ to $O(n^{-3/2})$, compared to the order $O(n^{-1})$ obtained from the conventional bootstrap approach advocated by Lee and Young (1999b). The same result holds for other forms of nonparametric likelihood, provided an expansion of the kind given by (13) of DiCiccio and Monti (2001) exists.

Remark 4. DiCiccio *et al.* (2001) show that in the parametric context, and for the specific case $u(Y, \psi) = \Phi\{r_p(\psi)\}$, the coverage error of the confidence set is reduced by weighted prepivoting to $O(n^{-3/2})$. This same order of error as that obtained from the analytic adjustment r_a to the signed root statistic r_p is achieved without any need for analytic calculation, or specification of the ancillary required by r_a. DiCiccio *et al.* (2001) argue that in this context weighted prepivoting is less effective when applied with other confidence set roots, such as those based on the Wald or score statistics. An empirical investigation of this claim for a particular inference problem is provided in Section 5 below.

Remark 5. We have assumed here that in the parametric context inference is required for the parameter of interest γ in the presence of the nuisance parameter ξ. We note that in the absence of any nuisance parameter, confidence sets based on the weighted prepivoted root $\tilde{u}_1(Y, \psi)$ will always have *exactly* the desired coverage $1 - \alpha$. The conventional bootstrap approach, based on $\widehat{u}_1(Y, \psi)$, will only yield exact inference if the initial root $u(Y, \psi)$ is exactly pivotal. There seem, therefore, strong arguments in favour of general adoption of weighted bootstrap schemes.

5. Numerical illustrations

Illustration 1: Exponential regression. Consider an exponential regression model in which $T_1, \ldots, T_n$ are independent, exponentially distributed lifetimes, with means of the form $E(T_i) = \exp(\beta + \xi z_i)$, with known covariates $z_1, \ldots, z_n$. Suppose that inference is required for the mean lifetime for covariate value z_0.

Let the parameter of interest therefore be $\gamma = \beta + \xi z_0$, with nuisance parameter ξ. The signed root likelihood ratio statistic is

$$r_{\mathrm{p}}(\gamma) = \mathrm{sgn}(\widehat{\gamma} - \gamma)\left[2n\left\{(\gamma - \widehat{\gamma}) + (\widehat{\xi}_\gamma - \widehat{\xi})\bar{c} + n^{-1}\exp(-\gamma)\sum_{i=1}^{n} T_i \exp(-\widehat{\xi}_\gamma c_i) - 1\right\}\right]^{1/2},$$

where $c_i = z_i - z_0$, $i = 1, \ldots, n$ and $\bar{c} = n^{-1}\sum c_i$.

In this case, the calculations leading to the adjusted version r_a of r_p are readily performed. However, it is easily verified that r_p is exactly pivotal. To see this, substitute $T_i = \exp(\gamma + \xi c_i)Y_i$, where the Y_i are independently, exponentially distributed with mean 1, and observe that the signed root statistic may be expressed as a (complicated) function of $Y_1, \ldots, Y_n$, and so has a distribution which does not depend on (γ, ξ). Therefore, even conventional bootstrapping yields the true sampling distribution, modulo simulation error. There is no need for weighted bootstrapping in this problem.

For numerical illustration, we consider data extracted from Example 6.3.2 of Lawless (1982), as analyzed by DiCiccio *et al.* (2001). The $n = 5$ responses T_i are 156, 108, 143, 56 and 1, survival times in weeks of patients suffering from leukaemia, and the corresponding covariate values are 2.88, 4.02, 3.85, 3.97 and 5.0, the base-10 logarithms of initial white blood cell count. We take $z_0 = \log_{10}(50,000)$. For these data, $\widehat{\gamma} = 2.399$ and $\widehat{\xi} = -2.364$.

We compare the coverage properties of confidence sets derived from $\Phi(r_p)$, $\Phi(r_a)$ and bootstrapping for $n = 5$, in an exponential regression model with these parameter values and the fixed covariate values. Table 1 compares actual and nominal coverages provided by the three constructions, based on 20,000 simulated datasets. Coverages based on normal approximation to r_p are quite inaccurate, but normal approximation to r_a provides much more accurate inference, while bootstrap confidence sets (each based on R=1999 bootstrap samples) display coverages very close to nominal levels.

TABLE 1. *Coverages (%) of confidence sets for mean* $\gamma = \exp(\beta + \xi z_0)$ *at* $z_0 = \log_{10}(50,000)$ *in exponential regression example, estimated from 20,000 data sets of size* $n = 5$ *and using* $R = 1,999$ *bootstrap replicates.*

Nominal	1.0	2.5	5.0	10.0	90.0	95.0	97.0	99.0
$\Phi(r_{\mathrm{p}})$	1.5	3.6	6.7	12.8	93.3	96.8	98.6	99.4
$\Phi(r_{\mathrm{a}})$	1.0	2.6	5.4	10.4	89.8	94.9	97.4	99.0
Bootstrap	1.0	2.5	5.1	10.0	89.9	94.8	97.4	98.9

Other cases where it is easily verified that r_p is exactly pivotal, and therefore conventional bootstrapping of r_p will provide exact inference, include inference for the error variance in a normal-theory linear regression model, and the related Neyman-Scott problem, as described by Barndorff-Nielsen and Cox (1994, Example 4.2).

Illustration 2: Normal distributions with common mean. We consider now the problem of parametric inference for the mean, based on a series of independent normal samples with the same mean but different variances. Initially we consider a version of the Behrens-Fisher problem in which we observe Y_{ij}, $i=1, 2$, $j=1, \ldots, n_i$, independent $N(\gamma, \sigma_i^2)$. The common mean γ is the parameter of interest, with orthogonal nuisance parameter $\xi=(\sigma_1, \sigma_2)$. Formally, this model is a (4,3) exponential family model. In such a model, the adjusted signed root statistic r_a is intractable, though readily computed approximations are available: see Skovgaard (1996); Severini (2000, Chapter 7).

We compare coverages of confidence sets derived from $\Phi(r_p), \Phi(\tilde{r}_a)$, the conventional bootstrap, which bootstraps at the overall maximum likelihood estimator $(\widehat{\gamma}, \widehat{\xi})$, and the weighted bootstrap, which uses bootstrapping at the constrained maximum likelihood estimator $(\gamma, \widehat{\xi}_\gamma)$, for 50,000 datasets from this model, with parameter values $\gamma=0$, $\sigma_1^2=1$, $\sigma_2^2=20$ and sample sizes $n_1=n_2=5$. All bootstrap confidence sets are again based on $R=1,999$ bootstrap samples. Also considered are the corresponding coverages obtained from $\Phi(W)$ and $\Phi(S)$ and their conventional and weighted bootstrap versions, where W and S are Wald and score statistics respectively, defined as the signed square roots of the statistics (3.33) and (3.35) of Barndorff-Nielsen and Cox (1994, Chapter 3). In the study, $\tilde{r}_a$ is an approximation to r_a based on orthogonal parameters (Severini, 2000, Chapter 7).

The coverage figures shown in Table 2 confirm that the simple bootstrap approach improves over asymptotic inference based on any of the statistics r_p, S, or W. Conventional bootstrapping yields very accurate inference for all three statistics: gains from using by the constrained bootstrap are slight. Overall, bootstrapping is very competitive in terms of accuracy when compared to $\tilde{r}_a$.

TABLE 2. *Coverages (%) of confidence intervals for Behrens-Fisher example, estimated from 50,000 data sets with bootstrap size* $R = 1,999$.

Nominal	1.0	2.5	5.0	10.0	90.0	95.0	97.5	99.0
$\Phi(r_p)$	1.6	3.7	6.5	11.9	88.0	93.5	96.4	98.4
MLE bootstrap	0.9	2.3	4.8	10.1	90.1	95.3	97.8	99.2
Constrained MLE bootstrap	0.8	2.3	4.8	10.0	90.1	95.3	97.9	99.2
$\Phi(\tilde{r}_a)$	0.9	2.5	5.2	10.6	89.5	94.9	97.6	99.1
$\Phi(W)$	5.2	8.0	11.3	16.5	83.5	88.9	92.2	94.8
MLE bootstrap	0.8	2.3	4.7	10.1	90.1	95.2	97.8	99.2
Constrained MLE bootstrap	0.8	2.2	4.6	10.0	90.2	95.4	97.9	99.3
$\Phi(S)$	0.1	1.4	4.7	11.6	88.6	95.5	98.7	99.9
MLE bootstrap	1.1	2.4	5.0	10.0	90.1	95.2	97.7	99.0
Constrained MLE bootstrap	0.9	2.4	5.0	10.1	90.0	95.2	97.7	99.1

In the above case, the nuisance parameter is two-dimensional. As a more challenging case, we consider extending the above analysis to inference on the common mean, set equal to 0, of six normal distributions, with unequal variances $(\sigma_1^2, \ldots, \sigma_6^2)$, set equal to $(1.32, 1.93, 2.22, 2.19, 1.95, 0.11)$, these figures being the variances for the data of Example 7.15 of Severini(2000, Chapter 7), which represent measurements of strengths of six samples of cotton yarn. The inference is based on an independent sample of size 5 from each population. Table 3 provides figures corresponding to those in Table 2 for this regime. Now the bootstrap approach is clearly more accurate than the approach based on $\tilde{r}_a$, and it is possible to discern advantages to the weighted bootstrap approach compared to the conventional bootstrap. Again, the weighted bootstrap works well when applied to the Wald and score statistics, casting some doubt on the practical significance of the arguments of DiCiccio *et al.* (2001).

In summary, it is our general experience that analytic approaches based on r_a are typically highly accurate when the dimensionality of the nuisance parameter is small and r_a itself is readily constructed, as in, say, a full exponential family model, where no ancillary statistic is required. In such circumstances, the argument for using the weighted bootstrap then rests primarily on its maintaining accuracy while avoiding cumbersome analytic derivations. In more complicated settings, in particular when the nuisance parameter is high dimensional or analytic adjustments r_a must be approximated, the weighted bootstrap approach is typically preferable both in terms of ease of implementation and accuracy. Gains over conventional bootstrapping may, however, be slight.

TABLE 3. *Coverages (%) of confidence intervals for normal mean example, estimated from 50,000 data sets with bootstrap size* $R = 1, 999$.

Nominal	1.0	2.5	5.0	10.0	90.0	95.0	97.5	99.0
$\Phi(r_{\mathrm{p}})$	3.0	5.7	9.3	15.1	85.3	91.2	94.6	97.2
MLE bootstrap	1.1	2.7	5.2	10.2	90.3	95.0	97.5	98.9
Constrained MLE bootstrap	0.9	2.5	5.1	10.1	90.4	95.2	97.6	99.0
$\Phi(\tilde{r}_{\mathrm{a}})$	1.5	3.4	6.4	11.9	88.7	93.9	96.7	98.5
$\Phi(W)$	6.7	9.6	13.3	18.7	82.0	87.3	90.8	93.6
MLE bootstrap	1.1	2.7	5.3	10.2	90.2	95.0	97.4	98.9
Constrained MLE bootstrap	0.9	2.4	5.0	9.9	90.5	95.3	97.6	99.1
$\Phi(S)$	0.6	2.1	5.1	10.9	89.6	95.2	98.0	99.5
MLE bootstrap	1.2	2.7	5.2	10.1	90.4	95.1	97.4	98.8
Constrained MLE bootstrap	1.1	2.5	5.2	10.2	90.3	95.1	97.5	99.0

Illustration 3: Nonparametric inference for variance. As a final illustration, we consider nonparametric inference for the variance $\gamma = 0.363$ of a folded standard normal distribution $|N(0,1)|$, for sample size $n = 50$.

From 20,000 datasets, we compared the coverage properties of confidence sets based on $u(Y,\psi) = \Phi\{(\widehat{\gamma}-\psi)/\widehat{\sigma}\}$, with $\widehat{\gamma}$ the sample variance and $\widehat{\sigma}^2$ an estimate of its asymptotic variance, and its conventional and weighted prepivoted forms $\widehat{u}_1(Y,\psi)$ and $\tilde{u}_1(Y,\psi)$. Table 4 displays the coverages of the three intervals. Weighted bootstrapping here utilised the exponentially tilted distribution involving empirical influence values described by Davison and Hinkley (1997, Section 9.4.1): see also DiCiccio and Romano (1990). Results for this, computationally simple, weighting procedure are very similar to those obtained from other, computationally less attractive, choices of construction of weighted bootstrap distribution.

Confidence sets based on $u(Y,\psi)$ are quite inaccurate, and substantial improvements are given by both conventional and weighted bootstrapping. Which of these is best depends, however, on the required coverage level. Similar conclusions are seen in other nonparametric examples: see Example 3 of Davison, Hinkley and Young (2003) and the examples given by Lee and Young (2003).

Graphical illustration of the prepivoting operation of the bootstrap is provided in Figure 1, which shows the distribution functions, as estimated from the 20,000 datasets, of $u(Y,\gamma)$, $\widehat{u}_1(Y,\gamma)$ and $\tilde{u}_1(Y,\gamma)$, with γ the true parameter value. The distribution of $u(Y,\gamma)$ is distinctly *not* Un$(0,1)$, while both bootstrap schemes yield prepivoted roots which *are* close to uniform, except in the lower tail. There, the distribution function of the conventional prepivoted root is closer to uniform than that of the weighted prepivoted root. The coverage figures shown in Table 4, of course, may be read directly off the graph of the distribution functions of the three confidence set roots.

TABLE 4. *Coverages (%) of bootstrap confidence sets for the variance γ when F is the folded standard normal distribution, estimated from 20,000 data sets of size $n = 50$ and using $R = 4,999$ bootstrap replicates; the root taken is $U(Y,\gamma) = \Phi\{(\widehat{\gamma}-\gamma)/\widehat{\sigma}\}$ with $\widehat{\gamma}$ sample variance.*

Nominal	1.0	2.5	5.0	10.0	90.0	95.0	97.5	99.0
$U(Y,\gamma)$	10.0	13.4	17.2	23.0	95.1	98.3	99.4	99.9
$\widehat{U}_1(Y,\gamma)$	3.1	5.4	8.5	14.1	91.6	96.5	98.6	99.6
$\tilde{U}_1(Y,\gamma)$	6.0	8.7	12.1	17.2	90.7	95.9	98.1	99.4

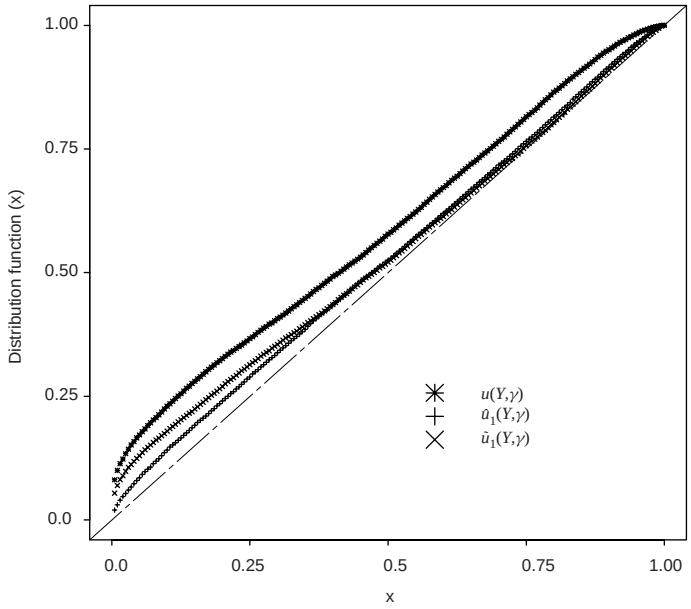

Figure 1.

6. Concluding remarks

In parametric inference, there are very persuasive arguments in favour of likelihood-based approaches to inference. In this context, all evidence points to weighted bootstrapping being an attractive alternative to analytic approaches, and yielding worthwhile improvements over conventional bootstrapping. In particular, the parametric illustrations presented here, and those considered by DiCiccio *et al.* (2001), demonstrate that excellent levels of accuracy may be obtained by the weighted bootstrap approach, which is easily implemented, without risk of impaired performance relative to conventional bootstrap methodology.

There is, of course, another consideration which should be taken into account in our analysis. Analytic approaches, such as those based on r_a, are designed to have conditional validity, given an ancillary statistic, as well as yield improved distributional approximation. There is some evidence that bootstrap approaches are less accurate from a conditional perspective. For the exponential regression example, exact conditional inference is described in Section 6.3.2 of Lawless (1982). Based on the same 5 observations as considered previously, Figure 2 provides graphical comparison between exact conditional significance levels, given ancillary statistics $\log T_i - \widehat{\gamma} - \widehat{\xi} c_i, i = 1, \ldots, n$, for testing the null hypothesis $\gamma = \widehat{\gamma} - \delta$, with approximate levels obtained from $\Phi(r_a)$ and

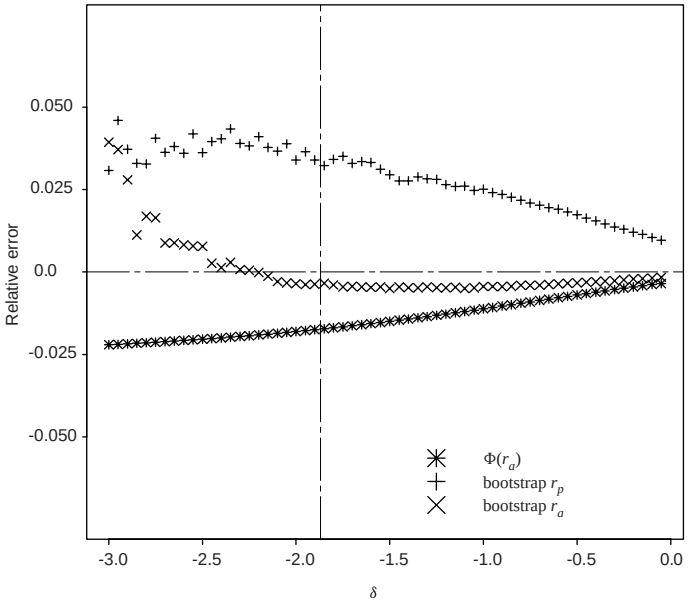

Figure 2.

by bootstrapping the distribution of r_p. Also shown are results obtained by bootstrapping the distribution of r_a. The figure plots the relative error of each approximation, defined as the approximation minus the true conditional significance level, expressed as a proportion of the true level, for a range of values of δ. Each bootstrap figure is based on $R = 10,000,000$ bootstrap samples. The conditional accuracy of the approximation obtained by unconditional bootstrapping of the distribution of r_p is less than that obtained from normal approximation to the distribution of the analytic adjustment r_a. However, bootstrapping the distribution of r_a gives excellent conditional accuracy, to very small significance levels. For reference, the vertical line shown in the figure corresponds to an exact conditional significance level of 1%. We conclude from this example and evidence presented by DiCiccio *et al.* (2001) that bootstrapping r_p provides satisfactorily stable inference, while bootstrapping r_a provides further conditional accuracy. In general, a statistical procedure is stable if it respects the principle of conditioning relative to any reasonable ancillary statistic, without requiring specification of the ancillary: see Barndorff-Nielsen and Cox (1994, Chapter 8). However, a full analysis of the stability properties of bootstrap inference is yet to be undertaken. One approach to conditional parametric bootstrapping in certain situations is through Metropolis-Hastings algorithms (Brazzale, 2000).

In the nonparametric context, the situation is less clear-cut. In particular, it is unclear whether the theoretical benefits of weighted bootstrapping over conventional bootstrapping are realisable in any particular situation, or whether weighted bootstrapping might actually reduce finite sample accuracy. Part of the problem here, of course, lies in the absence for many nonparametric problems of a 'gold standard', of the kind provided for the parametric setting by the adjusted signed root statistic r_a.

Our deliberations here therefore expose a number of issues for future development.

- There is a strong need to identify a gold standard for nonparametric inference. We might speculate, from evidence presented by Lee and Young (1999b), that this is provided by some form of (weighted) bootstrap calibration of nonparametric likelihood.
- We have attempted in this article to draw together various strands of contemporary bootstrap methodology. A systematic theory of weighted bootstrapping, encompassing both parametric and nonparametric and also likelihood and non-likelihood approaches to inference is suggested by our analysis and seems worth developing. That theory should encompass also two-sided inference for the parameter of interest.
- Our focus has been on the stylized problem of inference for a scalar parameter based on a random sample. Extensions to complex settings, in particular where the data do not constitute a random sample, seems desirable.

REFERENCES

BAGGERLY, K. A. (1998) Empirical likelihood as a goodness of fit measure, *Biometrika*, 85, 535–547.

BARNDORFF-NIELSEN, O. E. (1986) Inference on full or partial parameters based on the standardized signed log likelihood ratio, *Biometrika*, 73, 307–322.

BARNDORFF-NIELSEN, O. E. and COX, D. R. (1994) *Inference and Asymptotics*, London: Chapman & Hall.

BERAN, R. J. (1987) Prepivoting to reduce level error of confidence sets, *Biometrika*, 74, 457–468.

BERAN, R. J. (1988) Prepivoting test statistics: a bootstrap view of asymptotic refinements, *Journal of the American Statistical Association*, 83, 687–697.

BICKEL, P. J. and GHOSH, J. K. (1990) A decomposition for the likelihood ratio statistic and the Bartlett correction: a Bayesian argument, *Annals of Statistics*, 18, 1070–1090.

BRAZZALE, A. R. (2000) *Practical Small-Sample Parametric Inference*, Ph.D. thesis, Department of Mathematics, Swiss Federal Institute of Technology, Lausanne, Switzerland.

CARPENTER, J. (1999) Test inversion bootstrap confidence intervals, *Journal of the Royal Statistical Society series B*, 61, 159–172.

CORCORAN, S. A. (1998) Bartlett adjustment of empirical discrepancy statistics, *Biometrika*, 85, 967–972.

DAVISON, A. C. and HINKLEY, D. V. (1997) *Bootstrap Methods and Their Application*, Cambridge: Cambridge University Press.

DAVISON, A. C., HINKLEY, D. V. and YOUNG, G. A. (2003) Recent developments in bootstrap methodology, *Statistical Science*, 18, to appear.

DICICCIO, T. J., MARTIN, M. A. and STERN, S. E. (2001) Simple and accurate one-sided inference from signed roots of likelihood ratios, *Canadian Journal of Statistics*, 29, 67–76.

DICICCIO, T. J. and MONTI, A. C. (2001) Approximations to the profile empirical likelihood function for a scalar parameter in the context of $M-$estimation, *Biometrika*, 88, 337–351.

DICICCIO, T. J. and ROMANO, J. P. (1990) Nonparametric confidence limits by resampling methods and least favorable families, *International Statistical Review*, 58, 59–76.

DICICCIO, T. J. and ROMANO, J. P. (1995) On bootstrap procedures for second-order accurate confidence limits in parametric models, *Statistica Sinica*, 4, 141–160.

EFRON, B. (1979) Bootstrap methods: Another look at the jackknife, *Annals of Statistics*, 7, 1–26.

GARTHWAITE, P. H. and BUCKLAND, S. T. (1992) Generating Monte Carlo confidence intervals by the Robbins–Monro process, *Applied Statistics*, 41, 159–171.

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

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

HALL, P. and PRESNELL, B. (1999a) Intentionally biased bootstrap methods, *Journal of the Royal Statistical Society series B*, 61, 143–158.

HALL, P. and PRESNELL, B. (1999b) Biased bootstrap methods for reducing the effects of contamination, *Journal of the Royal Statistical Society series B*, 61, 661–680.

HALL, P. and PRESNELL, B. (1999c) Density estimation under constraints, *Journal of Computational and Graphical Statistics*, 8, 259–277.

LAWLESS, J. (1982) *Statistical Models and Methods for Lifetime Data*, New York: Wiley.

LEE, S. M. S. and YOUNG, G. A. (1999a) The effect of Monte Carlo approximation on coverage error of double-bootstrap confidence intervals, *Journal of the Royal Statistical Society series B*, 61, 353–366.

LEE, S. M. S. and YOUNG, G. A. (1999b) Nonparametric likelihood ratio confidence intervals, *Biometrika*, 86, 107–118.

LEE, S. M. S. and YOUNG, G. A. (2003) Prepivoting by weighted bootstrap iteration, *Biometrika*, 90, 393–410.

MARTIN, M. A. (1990) On bootstrap iteration for coverage correction in confidence intervals, *Journal of the American Statistical Association*, 85, 1105–1108.

OWEN, A. B. (1988) Empirical likelihood ratio confidence intervals for a single functional, *Biometrika*, 75, 237–249.

OWEN, A. B. (2001) *Empirical Likelihood*, Boca Raton: Chapman & Hall/CRC.

SKOVGAARD, I. M. (1996) An explicit large-deviation approximation to one-parameter tests, *Bernoulli*, 2, 145–165.

SEVERINI, T. A. (2000) *Likelihood Methods in Statistics*, Oxford: Clarendon Press.

G. ALASTAIR YOUNG
Statistical Laboratory
Centre for Mathematical Sciences
University of Cambridge
Wilberforce Road
Cambridge CB3 0WB (United Kingdom)
G.A.Young@statslab.cam.ac.uk