Ann. Inst. Statist. Math.
Vol. 52, No. 3, 507–518 (2000)

REDUCING BIAS WITHOUT PREJUDICING SIGN

PETER HALL[1], BRETT PRESNELL[1] AND BERWIN A. TURLACH[1,2]

[1] *Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia*
[2] *Co-operative Research Centre in Advanced Computational Systems, Australian National University, Canberra, ACT 0200, Australia*

(Received February 12, 1998; revised February 8, 1999)

Abstract. Jackknife and bootstrap bias corrections are based on a differencing argument which does not necessarily respect the sign of the true parameter value. Depending on sampling variability they can over-correct, producing a final estimator that is negative when one knows on physical grounds that it should be positive. To overcome this problem we suggest a simple, alternative bootstrap approach, based on biased-bootstrap methods. Our technique has similar properties to the standard uniform-bootstrap method in cases where the latter does not endanger sign, but it respects sign in a canonical way when the standard method disregards it.

Key words and phrases: Bias reduction, biased bootstrap, bootstrap, jackknife, twicing, weighted bootstrap.

1. Introduction

Methods for estimating bias were among the first applications of the bootstrap, appearing for example in Efron's (1979) seminal paper. They enjoy a close relationship to bias-correction techniques based on the jackknife, developed by Quenouille and Tukey thirty years prior to bootstrap methods. For example, the formula for a bias-reduced bootstrap estimator is similar to that for Tukey's ((1977), p. 526) approach based on twicing. An account of properties of bootstrap and jackknife estimators is given by Efron and Tibshirani ((1993), Chapter 10).

Twicing, and related jackknifed or bootstrapped bias-reduced estimators, are founded on cancelling out the dominant part of bias by subtracting an estimate of the mean of an estimator from twice its uncorrected value. If the estimator is highly variable, as can happen in small samples or in other cases where the relative error of the mean estimate is high, then this subtraction can produce a bias-reduced estimator which does not respect the sign that the true parameter value is known to have. For example, if a true mean is 0 then the standard bootstrap bias-reduced estimator of the square of the mean will be negative about 68% of the time. More generally, bootstrap bias-reduction can produce an estimator which violates the range of the parameter value. Efron's (1990) improved bias-reduced bootstrap estimator suffers from the same problems. The improvement that it offers is in terms of efficiency of the numerical algorithm used to compute it, and in fact it equals the standard bias-reduced bootstrap estimator if, when computing either, we conduct an infinite number of Monte Carlo resampling operations. Neither are transformations particularly helpful in reducing sign and range problems, since an unbiased estimator of a transformed parameter loses that virtue when back-transformed.

In this note we suggest a simple remedy for the sign and range problem, based

on the biased-bootstrap arguments of Hall and Presnell (1999). We shall focus attention on sign, which we feel is the most important case. Our approach guarantees that the bootstrap bias-reduced estimator shares the sign of the true parameter when the sign is known. It achieves this end by adjusting the bootstrap distribution on which the uncorrected estimator was based, rather than by adding a correction to that estimator. When the conventional, uniform-bootstrap estimator has the correct sign, its biased-bootstrap counterpart takes a similar value, and indeed has virtually identical large-sample properties in such cases. However, when the sign of the uniform-bootstrap estimator is incorrect, or only marginally correct, the biased-bootstrap estimator compensates by taking a value (with the correct sign) that is typically closer to the true parameter.

As in Efron (1990), our methods and theoretical arguments are developed for the case where the statistic of interest may be defined as a known function of a second statistic; it is the function that conveys the sign that we wish to preserve. Since we allow the argument of the known function to be vector-valued, this approach includes a wide range of settings, including statistics that are included in the general smooth function model considered by, for example, Hall ((1992), p. 52). Section 2 introduces our method and discusses its numerical properties. Section 3 describes its theoretical performance, and Section 4 sketches proofs of results in Section 3.

2. Methodology

2.1 *Uniform-bootstrap methods*

Let $\hat\theta$, a vector-valued statistic of length k, denote the uniform-bootstrap estimator of a quantity θ, based on data $\mathcal{X} = \{X_1, \ldots, X_n\}$. Suppose we wish to estimate $\psi^0 = \psi(\theta^0)$, where θ^0 is the true value of θ and ψ is a known, smooth function from $\mathbb{R}^k$ to $\mathbb{R}$. The bootstrap estimator is $\check\psi = \psi(\hat\theta)$, but is generally slightly biased. The standard uniform-bootstrap bias-reduced estimator is

$$\tilde\psi = 2\check\psi - E\{\psi(\hat\theta^*) \mid \mathcal{X}\}, \tag{2.1}$$

where $\hat\theta^*$ denotes the value of $\hat\theta$ computed from a resample $\mathcal{X}^* = \{X_1^*, \ldots, X_n^*\}$ drawn by sampling randomly and uniformly (that is, randomly with replacement) from $\mathcal{X}$. See for example Hall ((1992), p. 8ff), Efron and Tibshirani ((1993), p. 138) and Shao and Tu ((1995), p. 14f). The bias of $\tilde\psi$ is generally $O(n^{-2})$, compared with $O(n^{-1})$ in the case of $\check\psi$. However, this approach does not necessarily respect the sign of the function ψ. For example, if $\psi(u) \equiv u^2$, where θ is a population mean, then $\check\psi < 0$ is equivalent to $(n^{1/2}\bar X/S)^2 < 1$. Thus when $\theta^0 = 0$, the probability that $\check\psi < 0$ converges to the probability that a chi-squared random variable with one degree of freedom is less than 1; this limit equals 0.68.

2.2 *Biased-bootstrap bias correction*

Given a multinomial distribution $p = (p_1, \ldots, p_n)$ on the data $\mathcal{X} = \{X_1, \ldots, X_n\}$, let $\hat\theta^\dagger$ denote the version of $\hat\theta$ computed from a resample drawn by sampling at random from $\mathcal{X}$ according to the empirical distribution $\hat F_p$ that places mass p_i at X_i for $1 \le i \le n$. Let $\hat\theta_p$ denote the biased-bootstrap estimator of θ, or equivalently, the value that θ would assume if the true distribution were $\hat F_p$. (For example, if θ were the population mean then $\hat\theta_p$ would equal $\sum_i p_i X_i$.) Then, a biased-bootstrap bias-reduced estimator of ψ^0 is given by $\psi(\hat\theta_p) - \beta(p)$, where

$$\beta(p) = E_p\{\psi(\hat\theta^\dagger) \mid \mathcal{X}\} - \psi(\hat\theta) \tag{2.2}$$

and E_p denotes expectation in the biased-bootstrap distribution $\hat{F}_p$. In the event that $\beta(p) = 0$ we obtain simply the estimator $\psi(\hat{\theta}_p)$, which of course shares the sign of ψ. This suggests the following method. Let $p_{\text{unif}} = (n^{-1}, \ldots, n^{-1})$ be the uniform distribution on $\mathcal{X}$, and choose $p = \hat{p}$ to minimise the distance (see Subsection 2.3) between p and p_{unif} subject to $\beta(p) = 0$, $\sum_i p_i = 1$ and each $p_i \geq 0$. Then, our biased-bootstrap, bias-reduced estimator of ψ^0 is

$$\hat{\psi} = \psi(\hat{\theta}_{\hat{p}}), \tag{2.3}$$

where $\hat{\theta}_{\hat{p}}$ denotes the biased-bootstrap estimator of θ when $p = \hat{p}$.

Most importantly, if the function ψ is always of the one sign then $\hat{\psi}$ shares that sign, and in fact $\hat{\psi}$ is guaranteed to lie within the range of ψ. Moreover, $\hat{\psi}$ generally has asymptotic bias of order n^{-2}, compared with order n^{-1} for $\check{\psi}$; and each successive application of biased-bootstrap bias reduction reduces this order by the factor n^{-1}. Also, as in the case of conventional bootstrap bias reduction illustrated at (2.1), the asymptotic variance of $\hat{\psi}$ equals that of $\psi(\hat{\theta})$ in regular cases, and so the biased-bootstrap bias reduction method does not appreciably alter variance. Section 4 will address these properties in detail.

2.3 *Distance measures*

Appropriate distance functions include Cressie-Read (1984) power divergence distances. See Read and Cressie (1988) for a book-length discussion of power divergence. Specialised to the case of a multinomial distribution on n points, and for any given $-\infty < \rho < \infty$, the functional

$$D_\rho(p) = 2\{\rho(1-\rho)\}^{-1}\left\{n - \sum_{i=1}^{n}(np_i)^\rho\right\} \tag{2.4}$$

may be regarded as a measure of the distance (or divergence, since it is asymmetric) between p and p_{unif}. We define D_0 and D_1 by taking limits. Both are Kullback-Leibler distance measures, but it is conventional in problems of this type (see e.g. Owen (1988)) to take D_0, given by

$$D_0(p) = -2\sum_{i=1}^{n}\log(np_i)$$

(or any functional proportional to it), to be *the* measure of Kullback-Leibler distance in this setting.

2.4 *Numerical properties*

In problems where the basic statistic $\hat{\theta}$ can be written as a smooth function of an r-variate mean, computation may be undertaken using the method of estimating equations developed by Qin and Lawless (1994, 1995) for the case of empirical likelihood. If $\rho = 0$ then no modifications of Qin and Lawless' approach are required. In particular, Qin and Lawless' (1994) formula (3.2) for p_i is valid; for reference purposes we reproduce it here in their notation: $p_i = n^{-1}\{1 + t^T g(x_i, \theta)\}^{-1}$, where t is an r-vector of Lagrange multipliers and g is an r-variate function. When $0 < \rho < 1$ or $\rho = 1$, this should be altered to $p_i = \{c + t^T g(x_i, \theta)\}^{1/(1-\rho)}$ or $p_i = \exp\{c + t^T g(x_i, \theta)\}$, respectively, where the additional constant c, as well as t, is chosen so that the constraints $\sum_i p_i\, g(x_i, \theta) = 0$ and $\sum_i p_i = 1$ are satisfied.

Table 1. Panel (a) shows the percentages of times that the biased-bootstrap bias-reduced estimator failed to exist, and panel (b) gives the percentage of times that the uniform-bootstrap bias-reduced estimator was negative.

(a)

	$c = n^{1/2}\mu$			
n	0	1	2	5
25	6.7	3.3	1.1	0
50	3.5	2.2	0.7	0
100	1.7	0.6	0	0

(b)

	$c = n^{1/2}\mu$			
n	0	1	2	5
25	67.3	47.4	16.9	0
50	67.9	46.3	17.2	0
100	69.2	47.9	14.8	0

In the case of biased-bootstrap bias reduction, one of the estimating equations would generally enforce the constraint $\beta(p) = 0$. We then must optimise over the free parameters, either by solving simultaneously for c, t, and θ using a Newton-Raphson approach, or through a two-stage optimisation, solving for c and t (again Newton-Raphson is generally used here) for fixed values of θ and then minimising $D_\rho(p)$ over the free values of θ. As discussed in Section 5 of Qin and Lawless (1994), the first approach involves finding a saddle point of a function of c, t, and θ and may require particular care. For samples where the traditional uniform-bootstrap bias-reduced estimator does not respect sign, it is often the case that the value of $\hat{p}$ prescribed by the biased-bootstrap is quite far from p_{unif}. In such cases the simultaneous-optimisation approach may fail to converge when started from the uniform bootstrap estimates of θ (and $t = 0$). For particular samples this problem can be solved by using alternative starting values, chosen either by trial and error or by some more guided approach, but this may not be practical for simulation work. We have thus employed the two-stage approach in the simulation study reported here.

We consider the case of bias-correcting the square of the sample mean when the population mean is close to zero. In this case the constraint $\beta(p) = 0$ is equivalent to $(n-1)(\sum p_i X_i)^2 + \sum p_i X_i^2 = n\bar{X}^2$. Such a p exists if and only if the intersection of the convex hull of the points $\{(X_i, X_i^2) : i = 1, \ldots, n\}$ with the set $\{(x, y) : (n-1)x^2 + y^2 = \bar{X}^2\}$ is nonempty. From this it follows that $\hat{\psi}$ exists in the case of the sample $\mathcal{X}$ if and only if either all the X_i's have the same sign, or

$$\min\{-X_i : X_i < 0\} \times \min\{X_i : X_i > 0\} \leq n\bar{X}^2.$$

We also confine attention to the case $\rho = 0$. Simulation results for values of $\rho > 0$, specifically $\rho = 0.5$, showed no practically meaningful change in performance from those reported here, although a very small increase in bias could be detected.

We simulated from the $N(\mu = cn^{-1/2}, 1)$ distribution for sample sizes $n = 25$, 50 and 100, and $c = 0$, 1, 2 and 5. Panel (a) of Table 1 reports the percentages of times that the bias-bootstrap bias-reduced estimator $\hat{\psi}$ failed to exist. (Each entry in our tables is computed as the average of 1000 simulated values.) In each case where $\hat{\psi}$ did not exist, we replaced it by $\psi(\hat{\theta}) = \bar{X}^2$ when computing Monte Carlo approximations to bias and mean squared error. Panel (b) of the table gives the percentages of times that the uniform-bootstrap bias-reduced estimator $\tilde{\psi} = \bar{X}^2 - n^{-1}S^2$ (where $S^2 = n^{-1}\sum(X_i - \bar{X})^2$ denotes the sample variance) was negative. Negativity is seen to be a significant problem when the true mean is close to zero.

Tables 2 and 3 give Monte Carlo approximations to the biases and root mean squared errors of four different estimators of $\psi^0 = \psi(\theta^0)$: the unmodified estimator $\psi(\hat{\theta}) = \bar{X}^2$,

Table 2. Biases (shown as $n\times$ the value of the Monte Carlo approximation to bias) for the four estimators $\psi(\hat{\theta}) = \bar{X}^2$, $\tilde{\psi}$, $\tilde{\psi}_{\text{mod}}$ and $\hat{\psi}$, respectively.

c	n	X^2	U-Boot	Modif	B-Boot
0	25	0.99	0.03	0.69	0.50
1	50	0.03	0.06	0.72	0.53
	100	0.95	−0.03	0.65	0.46
1	25	0.92	−0.04	0.43	0.28
	50	0.98	0.00	0.46	0.30
	100	0.93	−0.06	0.42	0.26
2	25	0.84	−0.11	0.06	−0.00
	50	0.93	−0.05	0.13	0.06
	100	0.90	−0.08	0.07	0.01
5	25	0.63	−0.33	−0.33	−0.33
	50	0.78	−0.20	−0.20	−0.23
	100	0.83	−0.15	−0.15	−0.15

Table 3. Root mean squared errors (shown as $n\times$ the value of the Monte Carlo approximation to root mean squared error) for the four estimators $\psi(\hat{\theta}) = \bar{X}^2$, $\tilde{\psi}$, $\tilde{\psi}_{\text{mod}}$ and $\hat{\psi}$, respectively.

c	n	X^2	U-Boot	Modif	B-Boot
0	25	1.68	1.38	1.27	1.23
	50	1.83	1.51	1.42	1.38
	100	1.65	1.36	1.24	1.20
1	25	2.51	2.35	2.11	2.15
	50	2.58	2.39	2.16	2.21
	100	2.54	2.37	2.12	2.17
2	25	4.20	4.13	3.96	4.02
	50	4.27	4.17	3.99	4.06
	100	4.22	4.12	3.97	4.03
5	25	9.96	9.95	9.95	9.95
	50	10.13	10.10	10.10	10.10
	100	9.87	9.84	9.84	9.84

the uniform-bootstrap bias-reduced estimator $\tilde{\psi} = \bar{X}^2 - n^{-1}S^2$, the modified estimator $\tilde{\psi}_{\text{mod}}$ which equals $\tilde{\psi}$ if the latter is positive and equals $\psi(\hat{\theta})$ otherwise, and the biased-bootstrap bias-reduced estimator $\hat{\psi}$. Within each row of the tables these estimates are highly correlated, since all four estimators were computed for the same set of 1000 simulated samples. Thus within-row differences are estimated with very high accuracy. In Table 3, all of the within-row differences in estimated root mean square error are highly statistically significant except in the last three rows, corresponding to $c = 5$, where all differences might reasonably be attributed to simulation error. Similar remarks hold for the estimates of bias in Table 2. Here the apparent differences in bias are generally much larger than can be attributed to simulation error, although in several cases, and in particular when $c = 5$, the estimated biases of the various bias-corrected estimators are themselves not statistically significantly different from 0.

Table 2 shows that of all the estimators that are guaranteed to be positive, $\hat{\psi}$ has uniformly least bias. It is beaten by $\tilde{\psi}$ in the cases $c = 0, 1$, but there we know from

Table 1 (b) that $\tilde{\psi}$ suffers seriously from negativity problems. The biased-bootstrap estimator is always preferable when $c = 2$, surpassing the uniform-bootstrap estimator $\tilde{\psi}$ in terms of both bias and positivity. The two approaches are comparable in terms of bias when $c = 5$.

Table 3 gives Monte Carlo approximations to root mean squared errors for the four estimators addressed in Table 2. The biased-bootstrap estimator $\hat{\psi}$ is seen to improve on both $\psi(\hat{\theta})$ and $\tilde{\psi}$ in these terms. It is a little inferior to $\tilde{\psi}_{\text{mod}}$, although we know from Table 2 that it suffers less from bias than the latter. In cases where negativity is less of a problem (e.g. $c = 5$ in this simulation study), all four estimators perform virtually identically.

3. Theoretical properties

We consider the case where $\hat{\theta} = \bar{X} = n^{-1} \sum_i X_i$ is the mean of an r-variate sample $\mathcal{X}$, and ψ is a smooth function from r-dimensional Euclidean space to the real line. Examples include the case where ψ is a function of a ratio of two univariate means ($r = 2$), of a univariate variance ($r = 2$) or of a correlation coefficient ($r = 5$). We claim that in this setting, the biased bootstrap provides the same order of correction as the usual bootstrap.

To appreciate why, let $\mu^0 = E(X)$ denote the mean of the sampling distribution, where X is a generic X_i; let $X^{(j)}$ be the j-th component of X; and put

$$\psi_{j_1 \dots j_r}(x) = (\partial^r / \partial x^{(1)} \dots \partial x^{(j_r)}) \psi(x).$$

An elementary Taylor expansion argument shows that, provided the expected value of the remainder term may be bounded appropriately, the bootstrap estimator $\check{\psi} = \psi(\bar{X})$ has expected value $E\{\psi(\bar{X})\} = \psi(\mu^0) + \frac{1}{2} n^{-1} \xi + O(n^{-2})$, where $\xi = \sum_{j_1} \sum_{j_2} \operatorname{cov}(X^{(j_1)}, X^{(j_2)}) \psi_{j_1 j_2}(\mu^0)$. A bias-reduced estimator of $\psi(\mu^0)$ should provide a correction for the term $\frac{1}{2} n^{-1} \xi$ in this formula. The usual bootstrap bias-reduced estimator $\tilde{\psi}$, defined at (2.1), does this through having the property

$$\tilde{\psi} = \psi(\bar{X}) - \frac{1}{2} n^{-1} \xi + n^{-3/2} Y_n + O_p(n^{-2}), \tag{3.1}$$

where the random variable Y_n satisfies

$$E(Y_n) = 0 \quad \text{and} \quad E(Y_n^2) = O(1). \tag{3.2}$$

See for example Hall ((1992), p. 8ff). Theorem 3.1 below states that the same is true for the biased bootstrap estimator, $\hat{\psi}$, defined at (2.3).

Next we give our regularity conditions. Write $\|x\|$ for the Euclidean norm of a vector x. We assume that $E(\|X\|^4) < \infty$, that all fourth derivatives of ψ are uniformly bounded in some open neighbourhood of μ^0, and that the asymptotic variance of $n^{1/2} \psi(\bar{X})$, i.e.

$$\sigma^2 = \sum_{j_1} \sum_{j_2} \operatorname{cov}(X^{(j_1)}, X^{(j_2)}) \psi_{j_1}(\mu^0) \psi_{j_2}(\mu^0),$$

is nonzero. We call these conditions (A).

We specify the biased bootstrap estimator as follows. Let $C > 1$ be any constant, and let Π denote the set of vectors p such that $\sum_i p_i = 1$ and $0 \le p_i \le C n^{-1}$ for all

i. Let D_ρ be the distance function defined at (2.4), and let $\hat{\Pi} = \hat{\Pi}(\rho)$ denote the set of $p \in \Pi$ that give a local minimum of $D_\rho(p)$, subject to $\beta(p) = 0$ where β is defined at (2.2). Define $p = \hat{p} = \hat{p}(\rho)$ to be the element of $\hat{\Pi}$ that minimises $\|\bar{X} - \bar{X}_p\|$, if $\hat{\Pi}$ is nonempty, and put $\hat{p} = p_{\text{unif}}$ otherwise. Let $\hat{\psi}$ at (2.3) be defined using this $\hat{p}$.

THEOREM 3.1. *Assume conditions* (A), *and let* $0 \le \rho \le 1$. *Then*, (a) *with probability tending to* 1, $\hat{\Pi}$ *is nonempty, and* (b) (3.1) *continues to hold, for a random variable* Y_n *satisfying* (3.2), *if* $\tilde{\psi}$ *on the left-hand side is replaced by* $\hat{\psi}$.

One corollary of Theorem 3.1 is that, like the uniform-bootstrap estimator $\check{\psi} = \psi(\bar{X})$, the biased-bootstrap, bias-reduced estimator $\hat{\psi}$ is asymptotically normally distributed with mean $\psi(\mu^0)$ and variance $n^{-1}\sigma^2$. For general $\rho \in [0, 1]$, the versions of Y_n for $\tilde{\psi}$ and $\hat{\psi}$ may be taken to be identical, as we shall show in Section 4.

The case where ψ^0 is the square of a population mean is admitted by conditions (A), except when the true value of the mean is zero, since $\sigma^2 = 0$ in that instance. In fact, the theorem does not hold there; if it did then the biased-bootstrap, bias-reduced estimator $\hat{\psi}$ would share the shortcomings of its uniform-bootstrap counterpart. Our next result addresses this case.

We assume that the sample is univariate, and let S^2 denote its variance. Suppose $E(X) = 0$ and the index ρ of the distance function D_ρ satisfies $0 \le \rho < 1$. Given $y > 0$, define λ_0, λ_1, λ_2 (functions of y) by $E\{D(\lambda_0, \lambda_1, \lambda_2)\} = 1$, $E\{XD(\lambda_0, \lambda_1, \lambda_2)\} = 0$ and $E\{(X^2 - y)D(\lambda_0, \lambda_1, \lambda_2)\} = 0$, where $D(\lambda_0, \lambda_1, \lambda_2) = \{\lambda_0 + \lambda_1 X + \lambda_2(X^2 - y)\}^{-1/(1-\rho)}$. Put $d(y) = E\{D(\lambda_0, \lambda_1, \lambda_2)^\rho\}$ if $\rho \ne 0$, and $d(y) = -E\{\log D(\lambda_0, \lambda_1, \lambda_2)\}$ if $\rho = 0$; and given $0 < y < E(X^2)$, let $u = z_1(y)$ denote the point (assumed unique) at which the supremum of $d(u)$ over $0 < u \le y$ is achieved. (Since $d(0) = \infty$ then $z_1(y) > 0$.) Put $z(y) = y - z_1(y)$.

THEOREM 3.2. *Suppose* $\psi(u) \equiv u^2$, $\mu^0 = E(X) = 0$, $0 \le \rho < 1$, *and* $E(X^4) < \infty$. *Then for all* $\epsilon > 0$,

$$\hat{\psi} = \bar{X}^2 - n^{-1}S^2 + O_p(n^{-3/2}), \tag{3.3}$$

conditional on the event $n\bar{X}^2 - S^2 > \epsilon$; *and*

$$\hat{\psi} = n^{-1}z(n\bar{X}^2) + o_p(n^{-1}), \tag{3.4}$$

conditional on the event $n\bar{X}^2 - S^2 < -\epsilon$.

In the context of Theorem 3.2 the uniform-bootstrap, biased-reduced estimator of $\psi(\mu^0)$ is exactly $\tilde{\psi} = \bar{X}^2 - n^{-1}S^2$. In view of (3.3) the biased-bootstrap, bias-reduced estimator $\hat{\psi}$ agrees with $\tilde{\psi}$, up to smaller order terms, when $\tilde{\psi}$ is not troubled by sign problems. However, (3.4) shows that $\hat{\psi}$ departs significantly from $\tilde{\psi}$ when the latter has the wrong sign. It is generally the case that $z(y) = 0$ whenever $y < E(X^2)$, and there, (3.4) may be replaced simply by $\hat{\psi} = o_p(n^{-1})$. Hence, the biased-bootstrap correction not only enhances qualitative performance, by respecting sign, but can also improve the convergence rate in those cases where $\tilde{\psi}$ is negative. This was true for example in the simulation study reported in Section 2.4.

We could have stated (3.3) in the form (3.4), since conditional on $n\bar{X}^2 - S^2 > \epsilon$, $z_1(n\bar{X}^2) = \mu_2 = \sigma^2$ with probability tending to 1, and so $z(n\bar{X}^2) = n\bar{X}^2 - S^2 + o_p(1)$.

We chose the example of the square of the mean for its simplicity. One may readily develop versions of Theorem 3.2 in other cases, for example when ψ is a univariate, continuous, positive function with an isolated zero at μ^0. One may also treat the case where the population mean varies with n, for example in the form $n^{-1/2}c$ where $-\infty < c < \infty$. It may be proved in this general setting that the biased-bootstrap approach can give enhanced convergence rates when the uniform-bootstrap is troubled by sign problems, and performs similarly otherwise.

4. Technical arguments

4.1 *Outline proof of Theorem* 3.1

Denote components of $\bar{X}$ and X_i by superscripts in parentheses. Given $C_1 > 0$, let $\hat{P} = \hat{P}(C, C_1)$ be the set of vectors p such that $\sum_i p_i = 1$, $0 \le p_i \le Cn^{-1}$ for each i, and $\|\bar{X}_p - \bar{X}\| \le C_1 n^{-1}$. Define

$$\psi_{j_1 \dots j_r}(x) = (\partial^r / \partial x_{j_1} \cdots \partial x_{j_r}) \psi(x),$$

$$\hat{\sigma}_{j_1 \dots j_r}(p) = \sum_{i=1}^n (X_i^{(j_1)} - \bar{X}^{(j_1)}) \cdots (X_i^{(j_r)} - \bar{X}^{(j_r)}) p_i,$$

$$\hat{\tau}_{j_1 \dots j_r}(p) = \sum_{i=1}^n (X_i^{(j_1)} - \bar{X}_p^{(j_1)}) \cdots (X_i^{(j_r)} - \bar{X}_p^{(j_r)}) p_i,$$

$$\hat{\tau}^0_{j_1 \dots j_r}(p) = \sum_{i=1}^n |(X_i^{(j_1)} - \bar{X}_p^{(j_1)}) \cdots (X_i^{(j_r)} - \bar{X}_p^{(j_r)})| p_i,$$

$$\hat{T}_r(p) = \sum_{j_1} \cdots \sum_{j_r} \hat{\tau}_{j_1 \dots j_r}(p) \psi_{j_1 \dots j_r}(\bar{X}_p),$$

$$\hat{M}_r(p) = \max_{j_1 \dots j_r} |\hat{\tau}^0_{j_1 \dots j_r}|, \qquad \hat{M}(p) = \max\{\hat{M}_3(p), \hat{M}_4(p)\}.$$

Given $u > 0$, let $\langle u \rangle$ denote the smallest integer not strictly less than u. By Taylor expansion,

$$E_p\{\psi(\bar{X}^\dagger) \mid \mathcal{X}\} = \psi(\bar{X}_p) + \sum_{j=2}^3 (j!)^{-1} n^{-\langle j/2 \rangle} \hat{T}_j(p) + O_p\{n^{-2} \hat{M}_4(p)\}$$

$$= \psi(\bar{X}_p) + \frac{1}{2} n^{-1} \hat{T}_2(p) + O_p\{n^{-2} \hat{M}(p)\} \tag{4.1}$$

uniformly in $p \in \hat{P}$. Since $E(\|X\|^4) < \infty$ and each p_i is bounded by Cn^{-1} then $\hat{M}(p) = O_p(1)$ uniformly in $p \in \hat{P}$. And since additionally $\|\bar{X}_p - \bar{X}\|$ is bounded by $C_1 n^{-1}$ then $\psi_{j_1 j_2}(\bar{X}_p) = \psi_{j_1 j_2}(\bar{X}) + O_p(n^{-1})$, $\tau_{j_1 j_2}(p) = \hat{\sigma}_{j_1 j_2}(p) + O_p(n^{-1})$ and

$$\psi(\bar{X}_p) = \psi(\bar{X}) + \sum_j (\bar{X}_p - \bar{X})^{(j)} \psi_j(\bar{X}) + O_p(n^{-2}),$$

all uniformly in $p \in \hat{P}$. Combining the results from (4.1) down we deduce that, with $\beta(p) = E_p\{\psi(\bar{X}^\dagger) \mid \mathcal{X}\} - \psi(\bar{X})$ and

$$\hat{Q}(p) = \sum_j (\bar{X}_p - \bar{X})^{(j)} \psi_j(\bar{X}) + \frac{1}{2} n^{-1} \sum_{j_1} \sum_{j_2} \hat{\sigma}_{j_1 j_2}(p) \psi_{j_1 j_2}(\bar{X}),$$

we have

$$\sup_{p \in \hat{P}} |\beta(p) - \hat{Q}(p)| = O_p(n^{-2}). \tag{4.2}$$

Note that $\hat{Q}(p)$ is linear in p.

Given $C_2 > 0$, let $\{Z_n\}$ denote a sequence of random variables such that, for all n, $|Z_n| \le C_2$. Define $\hat{\sigma}_{j_1 j_2} = \hat{\tau}_{j_1 j_2}(p_{\text{unif}})$,

$$\hat{\sigma}^2 = \sum_{j_1} \sum_{j_2} \hat{\sigma}_{j_1 j_2} \psi_{j_1}(\bar{X}) \psi_{j_2}(\bar{X}), \qquad \hat{\xi} = \sum_{j_1} \sum_{j_2} \hat{\sigma}_{j_1 j_2} \psi_{j_1 j_2}(\bar{X}) = \hat{T}_2(p_{\text{unif}}).$$

Using the linearity of $\hat{Q}$ we may prove that the maximum of $\sum_i \log p_i$ subject to $\sum_i p_i = 1$ and $\hat{Q}(p) = n^{-2} Z_n$ occurs uniquely when $p = \tilde{p} = (\tilde{p}_1, \ldots, \tilde{p}_n)$, say, which admits the formula $\tilde{p}_i = n^{-1}\{1 + \lambda_{11} + \lambda_{12} \partial \hat{Q}(p)/\partial p_i\}^{-1/(1-\rho)}$ when $0 \le \rho < 1$, and $\tilde{p}_i = n^{-1} \exp\{1 + \lambda_{11} + \lambda_{12} \partial \hat{Q}(p)/\partial p_i\}$ when $\rho = 1$, where in each case the constants λ_{1j} are defined by the constraints $\sum_i \tilde{p}_i = 1$ and $\hat{Q}(\tilde{p}) = n^{-2} Z_n$. (The expressions for $\tilde{p}_i$ follow from a Lagrange multiplier argument.) Thus, for $0 \le \rho \le 1$ we have, by Taylor expansion, $\tilde{p}_i = n^{-1}\{1 + \lambda_{21} + \lambda_{22} \partial \hat{Q}(p)/\partial p_i + \cdots\}$, where the constants λ_{2j} are determined by the constraints, and the "$\cdots$" remainder term represents contributions to $\tilde{p}_i$ of powers of λ_{2j} higher than the first. Arguing in this manner we may prove that

$$\tilde{p}_i = n^{-1}\left[1 - \frac{1}{2} n^{-1} \hat{\xi} \hat{\sigma}^{-2} \sum_j (X_i - \bar{X})^{(j)} \psi_j(\bar{X}) + O_p\{n^{-2}(1 + \|X_i\|^3)\}\right], \tag{4.3}$$

where the remainder at (4.3) is of the stated order uniformly in i and in random sequences $\{Z_n\}$ such that $|Z_n| \le C_2$ for each n.

Let $\check{p} = (\check{p}_1, \ldots, \check{p}_n)$ denote any local maximiser of $\sum_i \log p_i$ subject to $p \in \hat{P}$ and $\beta(p) = 0$. By considering $n^{-2} Z_n$ to equal $-\hat{Q}(\check{p})$; and noting that for this Z_n, in view of (4.2),

$$\lim_{C_2 \to \infty} \limsup_{n \to \infty} P(|Z_n| > C_2) = 0;$$

we may, by choosing $C_2 = C_2(\epsilon)$ sufficiently large, deduce that for any given $\epsilon > 0$ the expansion (4.3) holds for $\check{p}$ as well as $\tilde{p}$, for each n, with probability greater than $1 - \epsilon$ (and with the same interpretation of the remainder as at (4.3)). Therefore,

$$\bar{X}_{\check{p}} = \bar{X} - \frac{1}{2} n^{-2} \hat{\xi} \hat{\sigma}^{-2} \sum_j (X_i - \bar{X})(X_i - \bar{X})^{(j)} \psi_j(\bar{X}) + O_p(n^{-2}), \tag{4.4}$$

$$\psi(\bar{X}_{\check{p}}) = \psi(\bar{X}) - \frac{1}{2} n^{-1} \hat{\xi} + O_p(n^{-2}). \tag{4.5}$$

From (4.4) we see that

$$\begin{aligned} \|\bar{X}_{\check{p}} - \bar{X}\| &= \{1 + o_p(1)\} \frac{1}{2} n^{-2} \hat{\xi} \hat{\sigma}^{-2} \left\| \sum_j (X_i - \bar{X})(X_i - \bar{X})^{(j)} \psi_j(\bar{X}) \right\| \\ &= \{1 + o_p(1)\} \frac{1}{2} n^{-1} \xi \sigma^{-2} \left\| \sum_j E\{(X - \mu^0)(X - \mu^0)^{(j)}\} \psi_j(\mu^0) \right\|. \end{aligned}$$

Hence, there exists a constant $C_3 > 0$ such that $P(\|\bar{X}_{\check{p}} - \bar{X}\| \le C_3) \to 1$ as $n \to \infty$. Therefore, since we define $\hat{p}$ to be the local maximum of $\sum_i \log p_i$ that is in $\hat{\Pi}$ (and subject to constraints) such that $\bar{X}_{\hat{p}}$ is *nearest* to $\bar{X}$, then provided we choose the constant C_1 (in the definition of $\hat{P}(C, C_1)$) greater than C_3, we shall have $\hat{p} = \check{p}$ with probability tending to 1 as $n \to \infty$. Part (a) of Theorem 3.1 follows from this fact. By (4.5),

$$\hat{\psi} = \psi(\bar{X}_{\hat{p}}) = \psi(\bar{X}) - \frac{1}{2} n^{-1} \hat{\xi} + O_p(n^{-2}). \tag{4.6}$$

Since $E(\|X\|^4) < \infty$ then by Taylor expansion, we may write $\hat{\xi} = \xi + n^{-1/2} Y_n + O_p(n^{-1})$ where the random variable Y_n satisfies (3.2). Part (b) of Theorem 3.1 follows from this property and (4.6).

Note too that by Taylor expansion,

$$\begin{aligned} E\{\psi(\bar{X}^*) \mid \mathcal{X}\} &= \psi(\bar{X}) + \sum_{j_1} \sum_{j_2} E\{(\bar{X}^* - \bar{X})^{(j_1)} (\bar{X}^* - \bar{X})^{(j_2)} \mid \mathcal{X}\} \psi_{j_1 j_2}(\bar{X}) \\ &\quad + O_p(n^{-2}) \\ &= \psi(\bar{X}) + \frac{1}{2} n^{-1} \hat{\xi} + O_p(n^{-2}), \end{aligned}$$

where $\bar{X}^*$ denotes the mean of a uniform-bootstrap resample. Hence, the uniform-bootstrap bias-reduced estimator $\tilde{\psi}$, given by (2.1), also admits the expansion (4.6). This proves that Y_n at (3.1) may be taken identical in the cases where the left-hand side is $\tilde{\psi}$ or $\hat{\psi}$.

4.2 *Outline proof of Theorem* 3.2

For the sake of brevity we consider only the case $\rho = 0$, corresponding to Kullback-Leibler loss. (There, the Lagrange multiplier λ_0 appearing in the argument of $D(\lambda_0, \lambda_1, \lambda_2)$ may be taken equal to 1.) Define $\hat{\mu}_2 = n^{-1} \sum_i X_i^2$, and observe that

$$E_p\{(\bar{X}^\dagger)^2 \mid \mathcal{X}\} = n^{-1} \sum_{i=1}^n p_i X_i^2 + (1 - n^{-1}) \bar{X}_p^2. \tag{4.7}$$

Let $\epsilon > 0$ be given, and consider first the case where $n\bar{X}^2 \ge \hat{\mu}_2$. Using a Lagrange multiplier argument we may show that the vector $p = \hat{p}$ that maximises $\sum_i \log p_i$ subject to $\sum_i p_i = 1$ and $\beta(p) = 0$ is given by

$$\hat{p}_i = n^{-1}[1 + \hat{\mu}_2^{-1}(t - \bar{X})(X_i - t) + O_p\{n^{-1}(1 + |X_i|^3)\}],$$

uniformly in i, where $t = \operatorname{sgn}(\bar{X})(\bar{X}^2 - n^{-1}\hat{\mu}_2)^{1/2}$. Hence,

$$\bar{X}_p = \bar{X} + \hat{\mu}_2^{-1}(t - \bar{X})(\hat{\mu}_2 - t\bar{X}) + O_p(n^{-1}) = t + O_p(n^{-1}),$$

which implies that

$$\hat{\psi} = t^2 + O_p(n^{-3/2}) = \bar{X}^2 - n^{-1}\hat{\mu}_2 + O_p(n^{-3/2}) = \bar{X}^2 - n^{-1}S^2 + O_p(n^{-3/2}),$$

as had to be shown.

Now suppose $n\bar{X}^2 \le \hat{\mu}_2 - \epsilon$. We begin by describing an algorithm that is equivalent to biased-bootstrap bias reduction in this setting. Let x and y denote candidates for the mean and mean square, respectively, of the biased-bootstrap distribution, and define

$$p_i = p_i(x,y) = n^{-1}\{1 + \lambda_1(X_i - x) + \lambda_2(X_i^2 - y)\}^{-1},$$

where λ_1, λ_2 are chosen as functions of x and y so that

$$\sum_{i=1}^n (X_i - x)p_i = \sum_{i=1}^n (X_i^2 - y)p_i = 0.$$

(These values of λ_1, λ_2 are of course stochastic, and so differ from those introduced just prior to the statement of Theorem 3.2. The latter will be denoted here by λ_1^0, λ_2^0.) Noting (4.7), express y as a function of x by $n^{-1}y + (1 - n^{-1})x^2 = \bar{X}^2$, i.e. $y = n\bar{X}^2 - (n-1)x^2$. This defines λ_1, λ_2, p_i as functions of x alone. Now choose $x = \hat{x}$ to maximise $\sum_i \log p_i$, or equivalently, to minimise $\hat{\alpha}\{x, y(x)\}$, where

$$\hat{\alpha}(x,y) = n^{-1}\sum_{i=1}^n \log\{1 + \lambda_1(x,y)(X_i - x) + \lambda_2(x,y)(X_i^2 - y)\}.$$

Then the biased-bootstrap estimator is $\hat{\psi} = \hat{x}^2$, and it may be proved that $\hat{x} \to 0$ in probability.

We approximate λ_1, λ_2 using a non-empirical version of this construction, as follows. Let X have the distribution of a generic X_i, let $\lambda_1^0 = \lambda_1^0(x,y)$ and $\lambda_2^0 = \lambda_2^0(x,y)$ be the functions determined by the equations

$$E[(X - x)\{1 + \lambda_1^0(X - x) + \lambda_2^0(X^2 - y)\}^{-1}] = 0,$$
$$E[(X^2 - y)\{1 + \lambda_1^0(X - x) + \lambda_2^0(X^2 - y)\}^{-1}] = 0,$$

and put

$$\alpha(x,y) = E[\log\{1 + \lambda_1^0(x,y)(X - x) + \lambda_2^0(x,y)(X^2 - y)\}].$$

Then, for any $0 < \epsilon_1 < \frac{1}{2}\mu_2$ there exists $\epsilon_2 > 0$ such that $\hat{\alpha}(x,y) = \alpha(x,y) + o_p(1)$ uniformly in pairs (x,y) such that $|x| \le \epsilon_2$ and $\epsilon_1 \le y \le \mu - \epsilon_1$. Moreover, ϵ_2 may be chosen so that $\alpha(x,y)$ is a uniformly continuous function of its argument on this set, and so

$$\hat{\alpha}\{\hat{x}, n\bar{X}^2 - (n-1)\hat{x}^2\} = \alpha(0, n\bar{X}^2 - n\hat{x}^2) + o_p(1).$$

It follows from this identity, and the fact that the supremum of $\alpha(0,u)$ over $0 < u \le n\bar{X}^2$ is achieved at $z_1(n\bar{X}^2)$, that if $n\hat{x}^2 - z(n\bar{X}^2)$ does not converge in probability to zero then we can produce a (stochastic) candidate $\tilde{x}$ for x such that

$$\hat{\alpha}\{\hat{x}, n\bar{X}^2 - (n-1)\hat{x}^2\} > \hat{\alpha}\{\tilde{x}, n\bar{X}^2 - (n-1)\tilde{x}^2\},$$

the latter inequality holding with probability bounded away from 0 along a subsequence. This would contradict the definition of $\hat{x}$ as the minimiser of $\hat{\alpha}\{x, y(x)\}$, and so $n\hat{x}^2 - z(n\bar{X}^2) \to 0$ in probability.

Acknowledgements

We are grateful to two referees for their helpful comments.

References

Cressie, N. A. C. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests, *J. Roy. Statist. Soc. Ser. B*, **46**, 440–464.

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

Efron, B. (1990). More efficient bootstrap computations, *J. Amer. Statist. Assoc.*, **82**, 171–200.

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

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

Hall, P. and Presnell, B. (1999). Intentionally-biased bootstrap methods, *J. Roy. Statist. Soc. Ser. B*, **61**, 143–158.

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

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations, *Ann. Statist.*, **22**, 300–325.

Qin, J. and Lawless, J. (1995). Estimating equations, empirical likelihood and constraints on parameters, *Canad. J. Statist.*, **23**, 145–159.

Read, T. R. C. and Cressie, N. A. C. (1988). *Goodness-of-fit Statistics for Discrete Multivariate Data*, Springer, New York.

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*, Springer, New York.

Tukey, J. W. (1977). *Exploratory Data Analysis*, Addison-Wesley, Reading, Massachusetts.