Applications of Intentionally Biased Bootstrap Methods

Peter Hall and Brett Presnell

Abstract. A class of weighted-bootstrap techniques, called biased-bootstrap methods, is proposed. It is motivated by the need to adjust more conventional, uniform-bootstrap methods in a surgical way, so as to alter some of their features while leaving others unchanged. Depending on the nature of the adjustment, the biased bootstrap can be used to reduce bias, or reduce variance, or render some characteristic equal to a predetermined quantity. More specifically, applications of bootstrap methods include hypothesis testing, variance stabilisation, both density estimation and nonparametric regression under constraints, 'robustification' of general statistical procedures, sensitivity analysis, generalised method of moments, shrinkage, and many more.

1991 Mathematics Subject Classification: Primary 62G09, Secondary 62G05
Keywords and Phrases: Bias reduction, empirical likelihood, hypothesis testing, local-linear smoothing, nonparametric curve estimation, variance stabilisation, weighted bootstrap

1. Uniform and weighted bootstrap methods

For centuries the sample mean has been recognised as an estimator of the population mean — or in contemporary notation, $\bar{X} = \int x\, d\widehat{F}(x)$ is an estimator of $\mu = \int x\, dF(x)$, where $\widehat{F}$ denotes the empirical distribution function computed using a sample drawn from a distribution F. The idea that the sample median is an estimator of the population median is implicit in work of Galton about 120 years ago. Thus, the notion that a parameter may be regarded as a functional of a distribution function, and estimated by the same functional of the standard empirical distribution, is a rather old one, even though it was perhaps only recognised as a general principle relatively recently.

Efron's (1979) classic paper on the bootstrap vaulted statistical science forward from these simple ideas. Efron saw that when substituting the true F by an estimator $\widehat{F}$, the notion of a 'parameter' could be interpreted much more widely than ever before. It could include endpoints of confidence intervals or critical points of hypothesis tests, as well as error rates of discrimination rules. It could encompass tuning parameters in a wide variety of estimation procedures (even the nominal levels of intervals or tests can be regarded as tuning parameters), and much more.

Another key ingredient of the methods discussed by Efron (1979) was recognition that in cases where the functional of $\widehat{F}$ could not be computed directly, it could be approximated to arbitrary accuracy by Monte Carlo methods. This differed in important respects from several earlier approaches to 'resampling', as the idea of sampling from the sample has come to be known. In particular, neither Mahalanobis' notion of 'interpenetrating samples', nor Hartigan's (1969) 'subsampling' approach, directly involve drawing a resample *of the same size as the original sample* by sampling with replacement. The methods of Simon (1969, Chapters 23–25) are closer in this respect to the contemporary bootstrap.

The combination of these two ideas — the substitution or 'plug in $\widehat{F}$' rule, and the notion that Monte Carlo methods can be used to surmount computational obstacles — has been little short of revolutionary. When Monte Carlo simulation is employed to compute a standard bootstrap estimator, one samples independently and uniformly from a data set $\mathcal{X} = \{X_1, \ldots, X_n\}$, producing a resample $\mathcal{X}^* = \{X_1^*, \ldots, X_n^*\}$ with the property that

$$P(X_i^* = X_j | \mathcal{X}) = n^{-1}\,, \quad 1 \leq i, j \leq n\,. \tag{1.1}$$

Standard bootstrap methods may be loosely defined as techniques that approximate the relationship between the sample and the population by that between the resample $\mathcal{X}^*$ and the sample $\mathcal{X}$.

The generality of the standard uniform bootstrap may be increased in a number of ways, for example by allowing the resampled values X_i^* to be exchangeable, rather than simply independent, conditional on $\mathcal{X}$ (see e.g. Mason and Newton, 1992); or by retaining the independence but replacing the sampling weight n^{-1} at (1.1) by p_j, say. In the latter case we shall use a dagger instead of the familiar asterisk notation, so that there will be no ambiguity about the procedure we are discussing:

$$P\big(X_i^\dagger = X_j \big| \mathcal{X}\big) = p_j\,, \quad 1 \leq i, j \leq n\,, \tag{1.2}$$

where $\sum_j p_j = 1$. This 'weighted bootstrap' procedure has been discussed extensively (see e.g. Barbe and Bertail, 1995), usually as a theoretical generalisation of the uniform bootstrap, pointing to a multitude of different modes of behaviour that may be achieved through relatively minor modification of the basic resampling idea.

2. Biased bootstrap methods

In 'standard' settings, where the appropriate way of applying the bootstrap is relatively clear, the uniform bootstrap offers an unambiguous approach to inference. Therein lies part of its attraction — there are no tuning parameters to be selected, for example. However, the lack of ambiguity can also be a drawback. In particular, the rigidity of the conventional bootstrap algorithm makes it relatively difficult to modify uniform-bootstrap methods so as to include constraints on the parameter space. The weighted bootstrap offers a way around this difficulty, by providing an opportunity for 'biasing' bootstrap estimators so as to fulfill constraints. Moreover, we may interpret the notion of a 'constraint' in a very broad sense, like that of a 'parameter'. Nevertheless, an unambiguous approach to choosing the weights p_i is required. Biased-bootstrap methods provide a solution to that problem.

The biased bootstrap requires two inputs from the experimenter: the distance measure, and the constraints. The first is generic to a wide range of problems, and will be discussed from that viewpoint in section 3. The second is problem-specific, and will be introduced through nine examples in section 4. A general form of the biased bootstrap is to choose the weights p_i so as to minimise distance from the distribution at (1.2) to that at (1.1), subject to the constraints being satisfied (Hall and Presnell, 1998a).

Details of some of the examples in section 4 may be found in Hall and Presnell (1998a,b,c) and Hall, Presnell and Turlach (1998). Examples not treated in section 4 include hypothesis testing, bagging (bootstrap aggregation), shrinkage, and applications involving time series data. The latter may be handled by either modelling the time series as a process with independent disturbances, and applying the biased bootstrap to those; or by using a biased form of the block bootstrap.

Section 5 will consider potential computational issues. Aids to computation include estimating equations, protected Newton-Raphson algorithms, and approximate, sequential linearisation. It will be clear that, using such techniques, biased-bootstrap methods are definitely computationally feasible.

3. DISTANCE MEASURES

For the sake of brevity we shall confine attention to a class of distance measures, the power divergence distances, introduced by Cressie and Read (1984) and Read and Cressie (1988). A wider range has been treated by Corcoran (1998) in the context of Bartlett adjustment of empirical likelihood. See also Baggerley (1998).

Let $p = (p_1, \ldots, p_n)$. For simplicity we assume throughout that $\sum_i p_i = 1$ and each $p_i \geq 0$, although in some cases (e.g. power divergence with index $\rho = 2$) the case where negative p_i's are allowed has computational advantages. Given $\rho \neq 0$ or 1, we may measure the distance between the uniform-bootstrap distribution, $p_{\text{unif}} = (n^{-1}, \ldots, n^{-1})$, and the biased-bootstrap distribution (with weight p_i at data value X_i) by

$$D_\rho(p) = \{\rho\,(1-\rho)\}^{-1} \left\{ n - \sum_{i=1}^{n} (np_i)^\rho \right\}.$$

This quantity is always nonnegative, and vanishes only when $p = p_{\text{unif}}$. For $\rho = \frac{1}{2}$, $D_\rho(p)$ is proportional to Hellinger distance. Letting $\rho \to 0$ we obtain

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

which equals half Owen's (1988) empirical log-likelihood ratio. Similarly, D_1 may be defined by a limiting argument; it is proportional to the Kullback–Leibler divergence between p and p_{unif} (whereas $D_0(p)$ is proportional to the Kullback–Leibler divergence between p_{unif} and p).

In constructing a biased-bootstrap estimator we would select a value of ρ, and then compute $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_n)$ from the sample $\mathcal{X} = \{X_1, \ldots, X_n\}$ so as to minimise $D_\rho(p)$, subject to the desired constraints being satisfied. If the parameter value that we wished to estimate was expressible as $\theta(F)$, then its biased-bootstrap estimator would equal $\theta(\widehat{F}_{\hat{p}})$, where $\widehat{F}_p$ denotes the distribution function of the discrete distribution that has mass p_i at data value X_i for $1 \le i \le n$. Usually the value of $\theta(\widehat{F}_{\hat{p}})$ will not be computable directly, but it may always be calculated by Monte Carlo methods, resampling from $\mathcal{X}$ according to the scheme that places weight $\hat{p}_i$ on X_i.

In some instances, for example outlier reduction (section 4.7), there are advantages to using $\rho \ne 0$, since $D_0(p)$ becomes infinite whenever some $p_i = 0$. By way of comparison, Hellinger distance (for example) allows one or more values of p_i to shrink to zero without imposing more than a finite penalty. However, in most other applications we have found that there is little to be gained — and sometimes, something to be lost (see sections 4.1 and 4.2) — by using a value of ρ other than $\rho = 0$.

4. Examples

4.1. Empirical likelihood. The method of empirical likelihood, or EL, was introduced by Owen (1988, 1990). See also Efron (1981). It may be viewed as a special case of the biased bootstrap in which the constraint is $\theta(\widehat{F}_p) = \theta_1$, where $\widehat{F}_p$ denotes the distribution function of the weighted bootstrap distribution with weights p_i, and θ_1 is a candidate value for θ. It is based on the value $\hat{p} = \hat{p}(\theta_1)$ of p that minimises $D_\rho(p)$ subject to $\theta(\widehat{F}_p) = \theta_1$.

One EL approach to constructing an α-level confidence interval for the true value of θ is to take t_α to be the upper α-level quantile of the chi-squared distribution for which the number of degrees of freedom equals the rank of the limiting covariance matrix of the uniform-bootstrap estimator, $\hat{\theta}(\widehat{F}_{p_{\mathrm{unif}}})$; and to let the interval be the set of θ_1's such that $D_\rho\{\theta(\widehat{F}_{\hat{p}(\theta_1)})\} \le t_\alpha$. Under regularity conditions that represent only a minor modification of those of Hall and La Scala (1990), this interval may be shown to have asymptotic coverage equal to $1 - \alpha$, no matter what the value of ρ. Using methods of DiCiccio, Hall and Romano (1991) it may be shown that this generalised form of EL is Bartlett-correctable if and only if $\rho = 0$. (Strictly speaking, the term 'likelihood' is appropriate for describing these generalised EL techniques only if $\rho = 0$.) See Baggerley (1998) and Corcoran (1998).

4.2. Variance stabilisation. Here we wish to choose, by empirical means, a transformation $\hat{g}$ which, when applied to a (scalar) parameter estimator $\hat{\theta}$, will implicitly correct for scale. Our method is a biased-bootstrap version of a conventional-bootstrap technique proposed by Tibshirani (1988). It has an advantage over the latter approach in that it does not require selection of any smoothing parameters, or any extrapolation.

As in example 4.1, choose p to minimise $D_\rho(p)$ subject to $\theta(\widehat{F}_p) = \theta_1$. Let $\mathcal{X}^\dagger = \{X_1^\dagger, \ldots, X_n^\dagger\}$ denote a resample drawn by sampling from $\mathcal{X}$ using the weighted bootstrap with weights $\hat{p}_i$, and let $\hat{\theta}^\dagger$ denote the version of $\hat{\theta}$ computed from $\mathcal{X}^\dagger$ rather than $\mathcal{X}$. Let $\hat{v}(\theta_1) = \text{var}(\hat{\theta}^\dagger|\mathcal{X})$ be the biased-bootstrap estimator of the variance of $\hat{\theta}$ when the true value of θ is θ_1. Write $\hat{g}(\theta)$ for the indefinite integral of $\hat{v}(\theta)^{-1/2}$, with the constant chosen arbitrarily. Using the uniform bootstrap, compute the conditional distribution of $\hat{g}(\hat{\theta}^*) - \hat{g}(\hat{\theta})$ and use it as an approximation to the unconditional distribution of $\hat{g}(\hat{\theta}) - \hat{g}(\theta^0)$, where θ^0 denotes the true parameter value. This enables us to compute confidence intervals for $\hat{g}(\theta^0)$, from which we may calculate intervals for θ^0 by back-transformation. It may be shown that $\rho = 0$ is sufficient for the latter intervals to be second-order accurate.

4.3. Density estimation under constraints. Here we consider kernel-type, biased-bootstrap estimators of the form $\hat{f}_p(x) = \sum_i p_i\, K_i(x)$, where $K_i(x) = h^{-1}K\{(x - X_i)/h\}$, K is a positive, symmetric kernel, and h is a bandwidth. (The traditional kernel estimator, in which each p_i is replaced by n^{-1}, may be regarded as a uniform-bootstrap estimator of $\theta = E\{K_i(x)\}$.) Constraining the j'th moment of the distribution with density $\hat{f}_p$ to equal the j'th sample moment is equivalent to asking that $\sum_i p_i\, A_i = a$, where a denotes the sample moment,

$$A_i = \sum_{k=0}^{\langle j/2 \rangle} \binom{j}{2k} X_i^{j-2k}\, h^{2k}\, \kappa_{2k}\,,$$

$\langle j/2 \rangle$ represents the integer part of $j/2$, and $\kappa_\ell = \int y^\ell\, K(y)\, dy$. Moreover, stipulating that the q'th quantile of the distribution with density $\hat{f}_p$ equal the q'th sample quantile ($\hat{\xi}_q$, say) produces a constraint of the same form, this time with $A_i = L\{(\hat{\xi}_q - X_i)/h\}$ (where L denotes the distribution function corresponding to the density K) and $a = q$. Constraining the interquartile range for $\hat{f}$ to equal its sample value amounts to the obvious linear form in constraints on the 25% and 75% quantiles. See also Chen (1997).

The constraint that entropy equals t, say, has the form

$$-\sum_{i=1}^{n} p_i \int K_i(x)\, \log\left\{\sum_{j=1}^{n} p_j\, K_j(x)\right\} dx = t\,.$$

Reducing entropy increases 'peakedness' and reduces spurious bumps in the tails. Combining this observation with the fact that increasing the bandwidth also tends to reduce the number of modes, while decreasing peakedness, we may develop an implicit algorithm (as distinct from the explicit method suggested in section 4.5) for computing a density estimator subject to the constraint of unimodality.

4.4. Correcting Nadaraya-Watson estimator for bias. Suppose data pairs (X_i, Y_i) are generated by the model $Y_i = g(X_i) + \epsilon_i$, where g is the smooth function that we wish to estimate, the design points X_i are random variables with density f, and the errors ϵ_i have zero mean. Then the Nadaraya–Watson estimator of g may be defined by $\tilde{g} = \widehat{\gamma}/\hat{f}$, where $\widehat{\gamma}(x) = n^{-1}\sum_i K_i(x)\, Y_i$ and $\hat{f}(x) = n^{-1}\sum_i K_i(x)$.

The performance of $\tilde{g}$ is generally inferior to that of local-linear estimators, owing to problems of bias. In particular, $\tilde{g}$ is biased for linear functions. To overcome this difficulty we may use the biased bootstrap to constrain the estimator to be unbiased when g is linear, by insisting that $\sum_i p_i(x)\,(x-X_i)\,K_i(x)=0$. Thus, $p=(p_1,\ldots,p_n)$ is now a function of location, x. The resulting estimator is

$$\hat g(x)=\left\{\sum_{i=1}^n p_i(x)\,K_i(x)\,Y_i\right\}\Big/\left\{\sum_{i=1}^n p_i(x)\,K_i(x)\right\}.$$

It achieves the same minimax efficiency bounds as local-linear smoothing (see e.g. Fan, 1993), and enjoys positivity properties that the latter approach does not.

4.5. Unimodality and monotonicity. Define a continuous density f to be strongly unimodal if there exist points $-\infty<x_1<x_2<\infty$ such that (i) f is convex on $(-\infty,x_1)$ and on (x_2,∞), and (ii) f is concave on (x_1,x_2). In principle we may constrain $\hat f_p$ to be a strongly unimodal density estimator, by arguing as follows: (a) for fixed x_1 and x_2, choose $p=p_{x_1x_2}$ to minimise $D_\rho(p)$ subject to $\hat f_p''(x)=\sum_i p_i\,K_i''(x)$ being positive on $(-\infty,x_1)$ and on (x_2,∞), and negative on (x_1,x_2); (b) choose x_1,x_2 to minimise $D_\rho(p_{x_1x_2})$ over all possible choices satisfying (a). However, the probability that this is possible does not necessarily converge to 1 as $n\to\infty$, even if the true f is strongly unimodal and considerable latitude is allowed for choice of bandwidth.

On the other hand, a weaker form of unimodality may be successfully imposed. There, we argue as follows: (α) select a candidate $-\infty<x_0<\infty$ for the mode of $\hat f_p$, and choose $p=p_{x_0}$ to minimise $D_\rho(p)$ subject to $\hat f'(x_0)=0$, $\hat f''(x_0)\le 0$, and to any point $x\ne x_0$ for which $\hat f'(x)=0$ being a point of inflexion of $\hat f_p$; and (β) choose x_0 to minimise $D_\rho(p_{x_0})$ over all possible choices satisfying (α). There is also a version of this method in the context of nonparametric regression, where 'unimodality' of a regression mean is defined in the obvious way.

Likewise, we may use biased-bootstrap methods to impose monotonicity of a function estimator in either the density or regression cases. Confining attention to local-linear estimators for nonparametric regression, we would proceed as follows. Let (X_i,Y_i), for $1\le i\le n$, denote a sample of independent and identically distributed data pairs. If (X_i,Y_i) is accorded weight p_i then the local-linear estimator of $g(x)=E(Y|X=x)$ equals $\hat a$, where $(\hat a,\hat b)$ denotes the pair (a,b) that minimises

$$\sum_{i=1}^n \{Y_i-a-b\,(X_i-x)\}^2\,p_i\,K_i(x)\,.$$

The biased-bootstrap local-linear estimator is $\hat g_p=(S_2T_0-S_1T_1)/(S_2S_0-S_1^2)$, where

$$S_j(x)=\sum_{i=1}^n (X_i-x)^j\,p_i\,K_i(x)\,,\quad T_j(x)=\sum_{i=1}^n Y_i\,(X_i-x)^j\,p_i\,K_i(x)\,.$$

Suppose we wish to constrain $\hat g_p(x)$ to have derivative not less than a given value t, for all x in some interval $\mathcal{I}$. It may be shown that, if the true regression mean g satisfies $g'\ge t$ on $\mathcal{I}$ then, with probability tending to 1 as $n\to\infty$, and for a wide range of choices of bandwidth, the biased-bootstrap constrained-minimisation problem has a solution, and that the solution has bias and variance of the same order as those of the unconstrained local-linear smoother.

4.6. Bias reduction without violating sign. Let $\hat{\theta} = \theta(\widehat{F}_{p_{\text{unif}}})$ (possibly vector-valued) denote the uniform-bootstrap estimator of $\theta(F)$, 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. The uniform-bootstrap estimator is $\check{\psi} = \psi(\hat{\theta})$, but is generally biased. The standard uniform-bootstrap bias-reduced estimator is $\widetilde{\psi} = 2\,\check{\psi} - E\{\psi(\hat{\theta}^*)|\mathcal{X}\}$, where $\hat{\theta}^*$ denotes the uniform-bootstrap version of $\hat{\theta}$. However, this approach does not necessarily respect the sign of the function ψ. For example, when $\psi(u) \equiv u^2$, and θ is a population mean and $\theta_0 = 0$, the probability that $\widetilde{\psi} < 0$ converges to 0.68.

A sign-respecting, biased-bootstrap approach to bias reduction may be defined as follows. 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 weighted empirical distribution $\widehat{F}_p$. A biased-bootstrap approximation to the bias of $\psi(\hat{\theta})$ is $\beta(p) = E_p\{\psi(\hat{\theta}^\dagger)|\mathcal{X}\} - \psi(\hat{\theta})$, where E_p denotes expectation with respect to $\widehat{F}_p$. Choose $p = \hat{p}$ to minimise $D_\rho(p)$ subject to $\beta(p) = 0$, $\sum_i p_i = 1$ and each $p_i \geq 0$. Then, our biased-bootstrap, bias-reduced, sign-respecting estimator of ψ_0 is $\widehat{\psi} = \psi(\hat{\theta}_{\hat{p}})$, where $\hat{\theta}_p = \theta(\widehat{F}_p)$.

It may be shown that, not only does $\widehat{\psi}$ overcome the sign problem, in cases where the probability that $\widetilde{\psi}$ has the wrong sign does not converge to 0, $\widehat{\psi}$ is closer (on average) than $\widetilde{\psi}$ to ψ_0.

4.7. 'Trimming' or 'winsorising.' Let $\hat{\theta}_p = \theta(\widehat{F}_p)$ denote the biased-bootstrap estimator of $\theta = \theta(F)$, and let $\gamma(p, \mathcal{X})$ be a measure of the concentration of the biased-bootstrap distribution with respect to $\hat{\theta}_p$. For example, in the case of a scalar sample $\mathcal{X}$, and when our interest is in location estimation, we might define

$$\gamma(p, \mathcal{X}) = \sum_{i=1}^{n} p_i \, (X_i - \bar{X}_p)^{2k} \,,$$

where $k \geq 1$ is an integer and $\bar{X}_p = \bar{X}_p(k)$ minimises $\sum_i \, p_i(X_i - x)^{2k}$ with respect to x. (Taking $k = 1$ we see that $\gamma(p, \mathcal{X})$ is the variance of the biased-bootstrap distribution.) Put $\widehat{\gamma} = \gamma(p_{\text{unif}}, \mathcal{X})$, being the version of the concentration measure in the case of the uniform bootstrap. Given $0 < t \leq \widehat{\gamma}$ we may calibrate the level of concentration by choosing $p = p(t)$ to minimise $D_\rho(p)$ subject to $\gamma(p, \mathcal{X}) = t$. As t decreases, the biased-bootstrap distribution $\widehat{F}_{p(t)}$ becomes more concentrated.

To avoid the result of calibration being heavily influenced by tail weight of the sampling distribution, we suggest 'inverting' the calibration so that it is on $D_\rho(p)$ rather than $\gamma(p, \mathcal{X})$. That is, given $\xi > 0$ we propose choosing $t = t_\xi$ such that $D_\rho\{p(t)\} = \xi$, and defining $\hat{p}(\xi) = p(t_\xi)$. In order for this approach to be practicable we require $D_\rho\{p(t)\}$ to be a monotone increasing function of t, which can be verified in many cases.

With this modification it may be shown that, in the case $0 < \rho \leq 1$, the biased bootstrap provides a remarkably effective device for reducing the effects of outlying data values. For example, in the context of univariate location estimation the estimator has a smooth, redescending influence curve, and a breakdown point that may be be located at any desired value $\epsilon \in (0, \frac{1}{2})$ simply by 'trimming' to a known distance (depending only on ϵ) from the empirical distribution. The estimator has an affine-equivariant multivariate form, and has versions for regression and nonparametric regression.

4.8. Sensitivity analysis. The ideas suggested in section 4.7 may be used to develop new, empirical methods for describing influence and sensitivity. For example, one may vary t by an infinitesimal amount, starting at $t = \widehat{\gamma}$, and rank the data values X_i in decreasing order of the amount by which this variation produces a *decrease* in the respective values of p_i. This may be regarded as ranking data values in terms of their influence on concentration, according to the chosen concentration measure. It produces an outlier diagnostic.

An alternative approach is to apply the biased bootstrap with θ equal to a candidate value, θ_1 say, for the parameter, and consider the values of $(\partial/\partial\theta_1)\, p_i(\theta_1)$ evaluated at the uniform-bootstrap estimator $\hat{\theta} = \theta(\widehat{F}_{\mathrm{unif}})$. (Of course, the signs of the derivatives convey important information about the nature of sensitivity.) Still another approach is to examine leave-one-out empirical-likelihood ratios computed at biased-bootstrap estimators. These influence diagnostics have potential advantages over traditional techniques; for example, they may be applied to quite arbitrary estimators and parameters.

4.9. Generalised method of moments. The generalised method of moments, or GMM, can provide substantial improvements over the naive method of moments, by reducing the variance of estimators. Versions of the biased bootstrap have already been successfully applied to GMM; see for example Brown and Newey (1995) and Imbens, Johnson and Spady (1998). However, those applications require equations defining the estimators to be of full rank, and the methods can perform poorly when one or more of those equations is (approximately) redundant. Indeed, one may show by example that in such cases, the rate of convergence of GMM estimators can be as slow as $n^{-1/4}$ (where n is sample size), rather than the $n^{-1/2}$ achieved using a much simpler method without a weight matrix in the least-squares step; and that this rate is not improved by iterating GMM. Biased-bootstrap methods can be used to identify redundancy and accommodate it adaptively. The approach involves choosing the weight matrix to minimise a non-asymptotic estimator of mean squared error, and thereby calibrating the standard GMM method so as to obtain nearly-optimal performance. The biased bootstrap is employed to enforce an empirical version of the method-of-moments constraint when defining the mean squared error estimator.

5. Computational issues

By way of notation, let us say that a constraint on p is linear if it may be written in the form $\sum_i p_i\, A_i = a$, which we denote by (L), where A_i and a depend only on the data, not on p, and may be vectors. (If they were vectors of length ν then we would, in effect, be imposing ν separate linear constraints.) Examples of linear constraints include those encountered in in the context of constraining moments and quantiles in section 4.3. Particularly for linear constraints, methods described by Owen (1990) and Qin and Lawless (1995), based on estimating equations, generally lead to numerically stable procedures.

It may be shown after a little algebra that under constraint (L), and when the distance function is D_ρ for some $\rho \neq 1$, the resulting p_i's are given by $p_i = p_i(\lambda) = (\lambda_0 + \lambda_1^T A_i)^{1/(1-\rho)}$, where λ_0 is a scalar, λ_1 is a column vector of length ν, and $\lambda = (\lambda_0, \lambda_1)$. (The λ_0 term comes from incorporating the additional condition $\sum_i p_i = 1$. We have not, at this stage, included the constraints $p_i \geq 0$, which in any event hold automatically when $-1 < \rho < 2$.) When $\rho = 1$ we have instead $p_i = \exp(\lambda_0 + \lambda_1^T A_i)$; and for any given ρ, the value of λ is defined by substituting back into (L). Thus, the dimension of the problem has been reduced from n to $\nu + 1$, which remains fixed as n increases. If in addition $\rho = 0$ then it may be shown that $\lambda_0 = n - \lambda_1^T a$, and so dimension reduces further, to ν.

In highly nonlinear problems, where these dimension reduction arguments do not apply, it may be necessary to compute the p_i's directly as the solution to an $(n-1)$-dimensional optimisation problem. For example, we have found that for moderate n a protected Newton-Raphson algorithm performs well in the problem of enforcing unimodality through constraints on entropy. Other approaches, such as the linearisation methods of Wood, Do and Broom (1996), may also be useful in nonlinear problems.

REFERENCES

Baggerley, K. A. (1998). Empirical likelihood as a goodness of fit measure. *Biometrika*, to appear.

Barbe, P. and Bertail, P. (1995). *The Weighted Bootstrap.* Springer, Berlin.

Brown, B. W. and Newey, W. K. (1995). Bootstrapping for GMM. Manuscript.

Chen, S. X. (1997). Empirical likelihood-based kernel density estimation. *Austral. J. Statist.* 39, 47–56.

Corcoran, S. A. (1998). Bartlett adjustment of empirical discrepancy statistics. *Biometrika*, to appear.

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

DiCiccio, T. J., Hall, P. and Romano, J. P. (1991). Empirical likelihood is Bartlett-correctable. *Ann. Statist.* 19, 1053–1061.

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

Efron, B. (1981). Nonparametric standard errors and confidence intervals. (With Discussion.) *Canad. J. Statist.* 36, 369–401.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.* 21, 196–216.

Hall, P. and La Scala, B. (1990). Methodology and algorithms of empirical likelihood. *Internat. Statist. Rev.* 58 109–127.

Hall, P. and Presnell, B. (1998a). Intentionally-biased bootstrap methods. *J. Roy. Statist. Soc. Ser. B*, to appear.

Hall, P. and Presnell, B. (1998b). Density estimation under constraints. Manuscript.

Hall, P. and Presnell, B. (1998c). Biased bootstrap methods for reducing the effects of contamination. Manuscript.

Hall, P., Presnell, B. and Turlach, B. (1998). Reducing bias without prejudicing sign. Manuscript.

Hartigan, J. A. (1969). Using subsample values as typical values. *J. Amer. Statist. Assoc.* 64, 1303–1317.

Imbens, G. W., Johnson, P. and Spady, R. H. (1998). Information theoretic approaches to inference in moment condition models. *Econometrica* 66, 333–358.

Mason, D. M. and Newton, M. A. (1992). A rank statistic approach to the consistency of a general bootstrap. *Ann. Statist.* 20, 1611–1624.

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

Owen, A. B. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.* 18, 90–120.

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.

Simon, J. L. (1969). *Basic Research Methods in Social Science.* Random House, New York.

Tibshirani, R. (1988). Variance stabilization and the bootstrap. *Biometrika* 75, 433–444.

Wood, A. T. A., Do, K.-A., and Broom, B. M. (1996). Sequential linearization of empirical likelihood constraints with application to U-statistics. *J. Computat. Graph. Statist.* 5, 365–385.

Peter Hall
Centre for Mathematics and its Applications
Australian National University
Canberra, ACT 0200
Australia
peter.hall@anu.edu.au

Brett Presnell
Centre for Mathematics and its Applications
Australian National University
Canberra, ACT 0200
Australia
brett.presnell@anu.edu.au