DATA-SMOOTHING AND BOOTSTRAP RESAMPLING

G.A. Young
Statistical Laboratory, University of Cambridge,
16 Mill Lane, Cambridge CB2 1SB, U.K.

1. INTRODUCTION

This paper reviews aspects of the smoothed bootstrap approach to statistical estimation.

The basic problem underlying the bootstrap methodology is that of providing a simulation algorithm which produces realisations from an unknown distribution F, when all that is available is a sample from F. The bootstrap of Efron (1979) simulates, with replacement, from the observed sample. The smoothed bootstrap, discussed by Efron (1979, 1982) and Silverman and Young (1987), smooths the sample observations first and hence effectively simulates from a kernel estimate of the density f underlying F. This is achieved, without construction of the kernel estimate itself, by resampling from the original data and then perturbing each sampled point appropriately.

The bootstrap and smoothed bootstrap will be considered as competing methods of estimating properties of an unknown distribution F. Given a general functional α, which may relate to the sampling properties of a parameter estimate, it is required to estimate on the basis of a set of sample data the population value $\alpha(F)$ of this functional.

The standard bootstrap estimates $\alpha(F)$ by $\alpha(F_n)$, F_n denoting the empirical c.d.f. of the sample data. The smoothed bootstrap estimates $\alpha(F)$ by $\alpha(\hat{F})$, where $\hat{F}$ is a smoothed version of F_n. The simple idea underlying the bootstrap estimation, therefore, is that of using F_n or $\hat{F}$ as a surrogate or estimate for the unknown F. In many circumstances the bootstrap estimate will itself be estimated by resampling from F_n or $\hat{F}$, though as yet unpublished work by Davison and Hinkley points in the direction of 'bootstrap resampling without the resampling'.

Though conceived by Efron (1979) as a means of tackling complex estimation problems, for a discussion of smoothing there is some advantage in studying the very simplest case where the functional α is linear in F. Relevant questions to be considered are:

(i) When is it advantageous to use a smoothed bootstrap rather than the standard bootstrap?
(ii) How should the smoothing be performed? Is there any advantage in simulating from a 'shrunk' version of the kernel estimator, with the same variance structure as the sample data?
(iii) Is it possible to define data-driven procedures which will choose the degree of smoothing to be applied automatically?

2. SMOOTHED BOOTSTRAP PROCEDURE

Suppose $X_1, \ldots, X_n$ are independent realisations from an unknown r-variate F. Assuming F has a smooth underlying density f, a convenient smoothed bootstrap is obtained from the kernel estimator $\hat{f}_{h,s}$ of f defined by

$$\hat{f}_{h,s}(x) = (1+h^2)^{r/2}\, \hat{f}_h\{(1+h^2)^{\frac{1}{2}} x\}\ ,$$
$$\hat{f}_h(x) = |V|^{-\frac{1}{2}}\, n^{-1}\, h^{-r} \sum_{i=1}^{n} K\{h^{-1}\, V^{-\frac{1}{2}}\, (x-X_i)\}\ . \tag{2.1}$$

Here K is a symmetric probability density function of an r-variate distribution with unit variance matrix. Operationally V is taken as the variance matrix of the sample data and h is a parameter defining the degree of smoothing.

Realisations generated from $\hat{f}_h$ have expectation equal to $\overline{X}$, the mean of the observed sample, but smoothing inflates the marginal variances. Silverman and Young (1987) give a number of simple examples which show that smoothing of this type can have a deleterious effect on the bootstrap estimation: see also section 3. The kernel estimator $\hat{f}_h$ is therefore 'shrunk' to give an estimator $\hat{f}_{h,s}$ with second-order moment properties the same as those in the observed sample. Note that the mean of $\hat{f}_{h,s}$ is $\overline{X}/(1+h^2)^{\frac{1}{2}}$.

3. LINEAR FUNCTIONALS

For a linear functional $\alpha(F) = \int a(t)dF(t)$, the smoothed bootstrap estimator is $\hat{\alpha}_h(F) = \int a(t)\hat{f}_{h,s}(t)dt$. This estimator may be written

$$\hat{\alpha}_h(F) = \frac{1}{n} \sum_{i=1}^{n} w^*(X_i) \tag{3.1}$$

where

$$w^*(x) = \int a\{(1+h^2)^{-\frac{1}{2}}\, (x+hV^{\frac{1}{2}}\xi)\}\, K(\xi)d\xi\ .$$

Using a Taylor expansion of a and the assumptions on the kernel function K, the mean squared error of $\hat{\alpha}_h(F)$ may, for h small, be expanded as

$$\text{MSE}\{\hat{\alpha}_h(F)\} = C_0 + C_1 h^2 + C_2 h^4 + O(h^6)\ . \tag{3.2}$$

Here we have assumed that $V = [V_{ij}]$ is a fixed positive definite symmetric matrix and

$$C_0 = \frac{1}{n} \int \{a(t) - \mu\}^2\, dF(t)\ ,$$

$$C_1 = \frac{1}{n} \int \{a(t) - \mu\}\, a^*(t)dF(t)\ ,$$

$$C_2 = \frac{1}{n} [\ 2 \int \{a(t)-\mu\}\ a^{**}(t)dF(t) + \frac{1}{4} \int a^*(t)^2 dF(t)$$

$$+ \frac{1}{4} (n-1) \{ \int a^*(t)dF(t) \}^2 \ ]$$

where $\mu = \int a(t)dF(t)$,

$$a^*(t) = D_V a(t) - t\cdot\nabla a(t) ,$$

$$a^{**}(t) = \frac{3}{8} t\cdot\nabla a(t) + \frac{1}{8} t^T H_a t - \frac{1}{2} D_V a(t) - \frac{1}{4} t\cdot\nabla(D_V a) + \frac{1}{8} D_V{}^2 a(t) .$$

Here $D_V a(t) = \sum_i \sum_j V_{ij} \delta^2 a(t)/\delta t_i \delta t_j$,

$$(H_a)_{ij} = \delta^2 a(t)/\delta t_i \delta t_j .$$

See Silverman and Young (1987) for details of the manipulations.

The expansion (3.2) immediately gives the result:

Lemma

Provided $a(X)$ and $a^*(X)$ are negatively correlated, the mean squared error of the smoothed bootstrap estimator $\hat{\alpha}_h(F)$ of $\alpha(F)$ will be less than that of the unsmoothed estimate $\hat{\alpha}_0(F) = \int a(t)dF_n(t)$, for *some* $h > 0$. □

The corresponding result for the bootstrap estimator $\tilde{\alpha}_h(F) = \int a(t)\hat{f}_h(t)dt$, constructed from the unshrunk kernel estimator, requires $a(X)$ and $D_V a(X)$ to be negatively correlated.

As a simple example, suppose F is the univariate standard Gaussian distribution and let $a(t) = t^5$. With $V = 1$ we have,

$$\text{cov}\{a(X), a^*(X)\} < 0$$

$$\text{cov}\{a(X), D_V a(X)\} > 0 ,$$

so that smoothing, with shrinkage, is of potential value in bootstrap estimation of the fifth moment.

The lemma above states that if $C_1 < 0$ in (3.2) some small degree of smoothing at least is worthwhile. If also $C_2 < 0$ we might speculate that some larger degree of smoothing may be appropriate. If both $C_1 > 0$ and $C_2 > 0$ the appropriate bootstrap estimator is the unsmoothed estimator $\hat{\alpha}_0(F)$. Otherwise, the optimal smoothing parameter, in the sense of minimising the approximate MSE $C_0 + C_1h^2 + C_2h^4$ is given by $h = (2|C_1|/4C_2)^{\frac{1}{2}}$.

The quantities C_1 and C_2 depend on the unknown underlying distribution function F , and in general will be complicated functions of the moments of F . A possible strategy would be to choose h with reference to a standard distribution, such as the standard r-variate Gaussian. In circumstances where the sample data do not suggest any sensible statistical model, C_1 and C_2 can be estimated, for example by substitution of the sample moments.

Given estimates $\hat{C}_1, \hat{C}_2$ for C_1, C_2 an entirely data-driven strategy for choosing the degree of smoothing would be to take $h = 0$ if $\hat{C}_1 \geq 0$, $h = \infty$ if $\hat{C}_1 < 0$ and $\hat{C}_2 < 0$ and $h = (2|\hat{C}_1|/4\hat{C}_2)^{\frac{1}{2}}$ otherwise. The case $h = \infty$ corresponds to Efron's 'parametric bootstrap' (Efron, 1979).

Rather than choosing h by reference to (3.2), which gives an expansion for h in the neighbourhood of zero, the representation (3.1) of the estimator can be used in conjunction with computer algebraic manipulation to obtain an exact expression for $\mathrm{MSE}\{\hat{\alpha}_h(F)\}$. This expression can then be minimised in h to obtain the optimal value of the smoothing parameter.

4. EXTENSION TO NON-LINEAR FUNCTIONALS

When an explicit bootstrap procedure is being used the functional α is unlikely to be linear. The ideas of Section 3 can be applied to bootstrap estimation for more general α , provided α admits a first-order von Mises expansion about F of the form

$$\alpha(\hat{F}) \simeq \alpha(F) + A(\hat{F} - F) , \qquad (4.1)$$

for $\hat{F}$ 'near' F . The functional α is linear and hence representable as an integral, $A(F) = \int a(t)dF(t)$, and to first-order the sampling properties of the bootstrap estimator $\alpha(\hat{F})$ of $\alpha(F)$ are the same as those of the estimator $A(\hat{F})$ of $A(F)$. Provided $\sup|\hat{F}-F|$ is $O_p(n^{-\frac{1}{2}})$, the error in (4.1) will be $O_p(n^{-1})$.

5. EXAMPLE

Let F be an unknown univariate distribution and consider estimation of the skewness,

$$\alpha(F) = \frac{E_F(X - E_F X)^3}{\{E_F(X-E_F X)^2\}^{3/2}} .$$

Simple manipulations, easily performed by computer algebra, show that the linear approximation (4.1) is defined by

$$a(t) = (t(-2\mu_1^4 t^2 + 3\mu_1^3\mu_2 t + 6\mu_1^3\mu_3 - 6\mu_1^2\mu_2^2 + 4\mu_1^2\mu_2 t^2 - 3\mu_1^2\mu_3 t$$
$$- 3\mu_1\mu_2^2 t - 6\mu_1\mu_2\mu_3 + 6\mu_2^3 - 2\mu_2^2 t^2 + 3\mu_2\mu_3 t))/2(\mu_1^6 - 3\mu_1^4\mu_2$$
$$+ 3\mu_1^2\mu_2^2 - \mu_2^3)\sqrt{(\mu_2 - \mu_1^2)} ,$$

where $\mu_r = E_F X^r$.

The bootstrap estimator $\hat{\alpha}_h(F)$ is given by:

$$\hat{\alpha}_h(F) = \frac{\sum_{i=1}^n X_i^3}{nV^{3/2}(1+h^2)^{3/2}} + \frac{3\bar{X}h^2}{V^{\frac{1}{2}}(1+h^2)^{3/2}} - \frac{3\bar{X}}{V^{\frac{1}{2}}(1+h^2)^{\frac{1}{2}}} - \frac{\bar{X}^3}{V^{3/2}(1+h^2)^{3/2}} . \quad (5.1)$$

In the special case of F standard Gaussian, computer algebraic manipulation of the function $a(t)$ gives a closed form approximation for the MSE of $\hat{\alpha}_h(F)$:

$$\text{MSE}\{\hat{\alpha}_h(F)\} \simeq \frac{6}{n(1+h^2)^3} , \quad (5.2)$$

and gives $C_1 = -18/n$, $C_2 = 36/n$. These values suggest, misleadingly, $h = \frac{1}{2}$.

In the general case, the formulae for C_1 and C_2 are complicated functions of the moments of F . With a manipulation package such as REDUCE it is straightforward to write FORTRAN subroutines to evaluate these coefficients: the moments of the observed sample are then substituted to yield estimates $\hat{C}_1, \hat{C}_2$. The formula for $\text{MSE}\{\hat{\alpha}_h(F)\}$, of which (5.2) is a special case, amounts to hundreds of lines of code. If $\mu_1 = 0$ it reduces to the simpler form:

$$\begin{aligned}\text{MSE}\{\hat{\alpha}_h(F)\} \simeq (&-8(h^2+1)^{\frac{1}{2}}h^4 n\mu_2^2\mu_3^2 - 16(h^2+1)^{\frac{1}{2}}h^2 n\mu_2^2\mu_3^2 \\ &+ 48(h^2+1)^{\frac{1}{2}}h^2\mu_2^2\mu_3^2 - 12(h^2+1)^{\frac{1}{2}}h^2\mu_2\mu_3\mu_5 - 8(h^2+1)^{\frac{1}{2}}n\mu_2^2\mu_3^2 \\ &+ 48(h^2+1)^{\frac{1}{2}}\mu_2^2\mu_3^2 - 12(h^2+1)^{\frac{1}{2}}\mu_2\mu_3\mu_5 + 4h^8 n\mu_2^2\mu_3^2 \\ &+ 16h^6 n\mu_2^2\mu_3^2 + 24h^4 n\mu_2^2\mu_3^2 - 9h^4\mu_2^2\mu_3^2 + 9h^4\mu_3^2\mu_4 \\ &+ 20h^2 n\mu_2^2\mu_3^2 + 36h^2\mu_2^5 - 24h^2\mu_2^3\mu_4 - 22h^2\mu_2^2\mu_3^2 \\ &+ 4h^2\mu_2^2\mu_6 + 18h^2\mu_3^2\mu_4 + 8n\mu_2^2\mu_3^2 + 36\mu_2^5 - 24\mu_2^3\mu_4 \\ &- 13\mu_2^2\mu_3^2 + 4\mu_2^2\mu_6 + 9\mu_3^2\mu_4) / (4n\mu_2^5(h^2+1)^4)) . \quad (5.3)\end{aligned}$$

Invariance of the estimator (5.1) under the transformation $X_i \to X_i+C$ $(i = 1,\ldots,n)$ suggests the following procedure for choice of h . Centre the observations X_i by calculating $Y_i = X_i - \bar{X}$ $(i = 1,\ldots,n)$. Then substitute $n^{-1}\sum_{i=1}^n Y_i^r$ for μ_r $(r = 2,\ldots,6)$ in (5.3). This gives an estimate of the mean squared error of the bootstrap estimator as a function of h . Use a numerical routine to minimise this and use the minimising value of h for the bootstrap estimation itself.

For each of four underlying distributions - standard Gaussian, uniform on [-1,1] , Beta (5,3) and standard exponential - and two sample sizes, n = 5 and n = 50 , 1000 datasets were generated. Table 1 shows, for each combination, the mean squared error over the 1000 replications of the bootstrap estimators $\hat{\alpha}_h(F)$, when h is chosen by various strategies. Strategy A takes h = 0.0 always, Strategy B takes h = 0.5 always, Strategy C estimates C_1, C_2 and chooses h according to the estimated values, as described in Section 3, while Strategy D is the procedure described above, based on (5.3).

Table 1 : MSE of bootstrap estimators, skewness example.

Distribution		N(0,1)	U[-1,1]	Beta(5,3)	Exp(1)
	$\alpha(F)$	0.0	0.0	-0.310	2.0
n	Smoothing Strategy				
5	A	0.3607	0.3566	0.3889	2.4497
	B	0.1847	0.1826	0.2341	2.7557
	C	0.2977	0.2950	0.3629	2.5674
	D	0.0912	0.0869	0.1554	3.0748
50	A	0.1092	0.0450	0.0650	0.4930
	B	0.0559	0.0230	0.0435	0.8661
	C	0.1066	0.0446	0.0649	0.5331
	D	0.0596	0.0218	0.0589	0.5490

The results of the simulation disappoint in that they do not provide concrete evidence in favour of any particular smoothing procedure. Automatic application of a small amount of smoothing can lead to substantially less accurate estimation: see the figure for the exponential simulation, n = 50 . Strategy C is unlikely to make the estimation dramatically worse and generally leads to some improvement over the standard bootstrap. Strategy D can lead to considerably greater accuracy in the bootstrap estimation but, as the exponential simulation makes clear, may also lead to quite inappropriate choice of h . Errors in the linear expansion (4.1), which is the basis of strategies C and D, may, even for moderate sample size, be quite appreciable.

Automatic procedures for choosing the degree of smoothing should be used with caution. It is probably advisable to examine the sample data, using an estimator of the form (2.1) say, and then to choose h with reference to some suggested parametric family of distributions.

Acknowledgement

I am grateful to Bernard Silverman for permission to include details of our joint work.

REFERENCES

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

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

Silverman, B.W. and Young, G.A. (1987). The bootstrap: to smooth or not to smooth? Biometrika, 74. (To appear)