BOOTSTRAPPING THE CORRELATION COEFFICIENT: A COMPARISON OF SMOOTHING STRATEGIES

DANIELA DE ANGELIS

Istituto di Statistica Economica, Università di Roma ‘La Sapienza’,
Piazzale A Moro 5, 00185 Roma, Italy

ALASTAIR YOUNG

Statistical Laboratory, University of Cambridge, 16 Mill Lane, Cambridge, U.K.

Estimation of the sampling standard deviation of the variance stabilised correlation coefficient by the smoothed bootstrap is reconsidered. Four different strategies for empirical choice of the smoothing parameter are discussed and their performances examined through a Monte Carlo experiment.

KEY WORDS: bootstrap, correlation coefficient, mean squared error, smoothed bootstrap, smoothing parameter.

1. INTRODUCTION

Use of the smoothed bootstrap in estimation of sampling properties of the correlation coefficient and its variance stabilised transform was introduced by Efron (1981, 1982). Silverman & Young (1987) reconsidered the idea, pointing out circumstances when smoothing the empirical distribution function may lead to reduction in mean squared error of the bootstrap estimator.

In Efron (1981, 1982), the parameter determining the amount by which the data are smoothed was fixed. In Silverman & Young (1987), though the question of an optimal choice of this parameter was considered, no obvious empirical strategy was suggested. The problem of choosing an appropriate smoothing parameter was discussed by Young (1988), where an empirical procedure based on minimisation of an estimate of the mean squared error of the smoothed bootstrap estimator was considered. Such a strategy, being very problem specific and involving precise assumptions, is not easily generalized to other contexts. A more general data-based procedure is essential if the smoothed bootstrap is to be routinely used as an alternative to the standard, unsmoothed, bootstrap.

In this paper estimation of the sampling standard deviation of the variance stabilised correlation coefficient is reconsidered. Four different strategies for an empirical choice of the smoothing parameter are discussed and their performances examined and compared through a small Monte Carlo experiment.
2. SMOOTHING THE BOOTSTRAP

The bootstrap estimation idea (Efron, 1979) is a very simple one. Let \( x_1, x_2, \ldots, x_n \) be a set of independent and identically distributed observations from an unknown distribution \( F \) and let \( \alpha(F) \) be some functional of interest. The bootstrap estimate of \( \alpha(F) \) is then \( \alpha(F_n) \), where \( F_n \) is the empirical distribution function, which places equal probability mass \( 1/n \) on each observed data point \( x_i \). Direct calculation of \( \alpha(F_n) \) is rarely feasible; \( \alpha(F_n) \) is more often computed by resampling from \( F_n \) i.e. with replacement from the observations \( x_1, x_2, \ldots, x_n \).

The smoothed bootstrap (Efron 1979, 1982; Silverman & Young 1987) is a modification of the original standard bootstrap which is particularly meaningful when the unknown \( F \) may be supposed continuous with density \( f \). The key step here is the replacement of \( F_n \), in both the estimation and the resampling phases, with an appropriate continuous version \( \hat{F}_h \) constructed by suitably smoothing \( F_n \). The smoothed bootstrap estimate of \( \alpha(F) \) is then simply \( \alpha(\hat{F}_h) \).

For the bootstrap application, a very convenient way of smoothing \( F_n \) is the following (Silverman & Young, 1987). Suppose \( F \) is a continuous \( d \)-variate distribution with density \( f \) and let \( x_1, x_2, \ldots, x_n \) be a set of observed values from \( F \). Then

\[
\hat{f}_h(x) = |V|^{-\frac{1}{2}} n^{-\frac{1}{2}} h^{-d} \sum_{i=1}^{n} K(h^{-1}V^{-\frac{1}{2}}(x - x_i))
\]

(1)

denotes the kernel method estimate (Parzen, 1962) of \( f \) at \( x \). Here \( V = [V_{ij}] \) represents the variance matrix of \( f \) or a consistent estimate of it and \( K(x) \) is a radially symmetric density function of a \( d \)-variate distribution with unit variance matrix. Integration of (1) gives the corresponding smoothed distribution function \( \hat{F}_h \). The amount of smoothing applied to the data is controlled by the single parameter \( h \), to be specified. The limiting case \( h = 0 \) gives the standard unsmoothed bootstrap.

It is important to note that evaluation of the smoothed bootstrap estimator \( \alpha(\hat{F}_h) \) does not require explicit construction of \( \hat{F}_h \). Sampling from \( \hat{F}_h \) may be performed by noting that if

\[
Y = x_I + h \varepsilon,
\]

where \( I \) is uniformly distributed on \( \{1, \ldots, n\} \) and \( \varepsilon \) has density \( K \), then \( Y \) has distribution \( \hat{F}_h \).

When \( \alpha(F) \) is a linear functional, of the form \( \alpha(F) = \int a(t) dF(t) \), it is possible to show (Silverman & Young, 1987) that under suitable differentiability properties on \( a \)

\[
MSE\{\alpha(\hat{F}_h)\} = MSE\{\alpha(F_n)\} + C_1 h^2 / n + O(h^4 + h^4 / n),
\]

(2)
as $h \to 0$ and $n \to \infty$, where $\text{MSE}\{\alpha(\hat{F})\}$ denotes the mean squared error of the estimator $\alpha(\hat{F})$ of $\alpha(F)$, and $C_1 = \text{cov}_F\{a(X), D_\gamma a(X)\}$, with $D_\gamma$ the operator defined by $D_\gamma a = \sum_i \sum_j V_{ij} \partial^2 a / \partial x_i \partial x_j$. It is clear from (2) that, provided $C_1 < 0$, taking $h > 0$ but to be $o(n^{-1})$ will lead to a reduction in the mean squared error of the bootstrap estimator. This conclusion can be extended to more general differentiable functionals $\alpha(F)$ by noting that if $\alpha(F)$ admits a first order von Mises expansion, for $\tilde{F}$ such that $\sup|\tilde{F} - F| = O_p(n^{-1})$, then

$$\alpha(\tilde{F}) = \alpha(F) + A(\tilde{F} - F) + O_p(n^{-1}) \tag{3}$$

with $A$ being linear. To a first level of approximation the sampling properties of $\alpha(\tilde{F})$ as an estimator of $\alpha(F)$ are the same as those of $A(\tilde{F})$ as an estimator of $A(F)$.

3. THE TRANSFORMED SAMPLE CORRELATION COEFFICIENT

An example of a non-linear functional is $\alpha(F) = (\text{var}_F Z)^{1/2}$, where $Z = \tanh^{-1} r$ with $r$ the sample correlation coefficient computed from $n$ independent and identically distributed observations from a bivariate distribution $F$.

For this estimation problem and the case of a Gaussian underlying distribution, Efron (1981, 1982) showed empirically that $\text{MSE}\{\alpha(\hat{F}_h)\} < \text{MSE}\{\alpha(\hat{F}_n)\}$ for certain $h > 0$. The same result was analytically obtained by Silverman & Young (1987) for a wider class of underlying distributions, using the approximation (3).

One question is not clear: in the references above the smoothing parameter was specified arbitrarily. It is only natural to ask whether a data-based strategy for an ‘optimal’ choice of $h$ might be devised.

4. EMPIRICAL SMOOTHING STRATEGIES

Four different data-based smoothing strategies have been studied in a Monte Carlo experiment. The results are summarised in Table 1. The experiment consisted of generating 100 samples of each of three sizes ($n = 10, 20, 50$) from each of four bivariate distributions: Gaussian (abbreviated as $N$ in Table 1), lognormal (LG) and Student’s $t$ with 3 and 10 degrees of freedom (T3 and T10 respectively). The lognormal and $t$ samples were generated through obvious transformations from a bivariate Gaussian distribution with mean zero, unit marginal variances and correlation coefficient $\rho = 0.5$. The smoothing parameter values produced by the four methods described below were employed in the construction of four different smoothed bootstrap estimators of $\alpha(F)$ from each of the 100 samples, for each combination of distribution and sample size. Table 1 shows the estimated mean squared errors obtained from the simulation of the bootstrap estimators corresponding to the different smoothing strategies.
Table 1  Estimated mean squared errors of the bootstrap estimators

<table>
<thead>
<tr>
<th></th>
<th>( \text{LSCV} )</th>
<th>( \text{BE} )</th>
<th>( \text{PLUG} )</th>
<th>( \text{QUICK} )</th>
<th>( \text{UNS} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.003599</td>
<td>0.007075</td>
<td>0.005893</td>
<td>0.005615</td>
<td>0.015726</td>
</tr>
<tr>
<td>20</td>
<td>0.001560</td>
<td>0.001558</td>
<td>0.002013</td>
<td>0.001907</td>
<td>0.002296</td>
</tr>
<tr>
<td>50</td>
<td>0.000327</td>
<td>0.000403</td>
<td>0.000420</td>
<td>0.000457</td>
<td>0.000721</td>
</tr>
<tr>
<td>10</td>
<td>0.00962</td>
<td>0.016234</td>
<td>0.008106</td>
<td>0.008646</td>
<td>0.020861</td>
</tr>
<tr>
<td>20</td>
<td>0.002817</td>
<td>0.004305</td>
<td>0.004376</td>
<td>0.004402</td>
<td>0.005121</td>
</tr>
<tr>
<td>50</td>
<td>0.000997</td>
<td>0.001232</td>
<td>0.001234</td>
<td>0.001240</td>
<td>0.001208</td>
</tr>
<tr>
<td>10</td>
<td>0.022688</td>
<td>0.034023</td>
<td>0.023842</td>
<td>0.025515</td>
<td>0.033267</td>
</tr>
<tr>
<td>20</td>
<td>0.017539</td>
<td>0.019539</td>
<td>0.018029</td>
<td>0.018142</td>
<td>0.018487</td>
</tr>
<tr>
<td>50</td>
<td>0.013560</td>
<td>0.014674</td>
<td>0.013717</td>
<td>0.013546</td>
<td>0.013244</td>
</tr>
<tr>
<td>10</td>
<td>0.002411</td>
<td>0.031055</td>
<td>0.022910</td>
<td>0.025104</td>
<td>0.026677</td>
</tr>
<tr>
<td>20</td>
<td>0.012863</td>
<td>0.014300</td>
<td>0.015187</td>
<td>0.015953</td>
<td>0.012042</td>
</tr>
<tr>
<td>50</td>
<td>0.016991</td>
<td>0.008097</td>
<td>0.008216</td>
<td>0.008403</td>
<td>0.007066</td>
</tr>
</tbody>
</table>

Each figure is based on 100 simulations.
Each bootstrap estimator is constructed using 50 resamples.

4.1 The LSCV Method

The first smoothing approach derives directly from (1) and consists of choosing \( h \) to obtain an accurate description of the unknown density \( f \). This is the notion underlying the least squares cross-validation (LSCV) method for selection of \( h \) in the nonparametric density estimation context (Bowman, 1984). This method is based on the minimisation of the score function

\[
CV(h) = \int \hat{f}_h(x) dx - 2n^{-1} \sum_{i=1}^{n} \hat{f}_{h,i}(x_i),
\]

where \( \hat{f}_{h,i} \) denotes the kernel estimate of \( f \) computed from the sample obtained from the original one by deleting the \( i \)th observation.

To obtain an easily computed approximation to \( CV(h) \), suppose that in (1) \( V \) is fixed. Then it is readily shown that

\[
\int \hat{f}_h(x) dx = n^{-2}h^{-2}|V|^{-\frac{1}{2}} \sum_i \sum_j K^{(2)}(h^{-1}V^{-1}(x_i - x_j)),
\]

where \( K^{(2)} \) denotes the convolution of \( K \) with itself. Also

\[
n^{-1} \sum_i \hat{f}_{h,i}(x_i) = n^{-1}(n - 1)^{-1}h^{-2}|V|^{-\frac{1}{2}} \sum_i \sum_j K(h^{-1}V^{-1}(x_i - x_j)) - (n - 1)^{-1}h^{-2}|V|^{-\frac{1}{2}}K(0).
\]

(4)
Replacing the factors of \((n - 1)\) in (4) by \(n\), we see that \(CV(h)\) may be approximated by

\[
ACV(h) = n^{-2}h^{-2}|V|^{-1} \sum_i \sum_j \{K^{(2)}[h^{-1}V^{-1}(x_i - x_j)] - 2K[h^{-1}V^{-1}(x_i - x_j)]\} + 2n^{-1}h^{-2}|V|^{-1}K(0).
\]  

(5)

The rationale behind the LSCV method becomes clear on observing that

\[
E_{\tilde{r}}[CV(h)] = MISE - \int f^2(x)dx
\]  

(6)

where the mean integrated squared error \(MISE\) (Rosenblatt, 1971) is a global measure of the quality of \(\hat{f}_h\) as an estimator of \(f\). From (6), it is immediate that minimisation of the cross-validation score function \(CV(h)\) is equivalent to minimisation of the unbiased estimate \(CV(h) + \int f^2(x)dx\) of the mean integrated squared error. Optimality of the LSCV strategy in terms of mean integrated squared error has been studied by Hall (1983) and Hall & Marron (1987).

In the case of a Gaussian kernel, (5) is particularly simple. Then

\[
ACV(h) \propto n^{-2}h^{-2} \sum_i \sum_j \{\frac{1}{4} \exp(-\frac{1}{4}h^{-2}d_{ij}) - 2 \exp(-\frac{1}{2}h^{-2}d_{ij})\} + 2n^{-1}h^{-2},
\]

where, writing \(x_i = (x_{i1}, x_{i2})\),

\[
d_{ij} = (x_{i1} - x_{j1})^2V^{11} + (x_{i2} - x_{j2})^2V^{22} + 2(x_{i1} - x_{j1})(x_{i2} - x_{j2})V^{12},
\]

with \(V^{-1} = [V^{ij}]\). In practice, the method is applied with \(V\) taken as the variance matrix of the sample. The NAG subroutine E04ABF was used to perform a minimisation of \(ACV(h)\) for each of the 100 samples generated for each combination of distribution and sample size. Mean squared errors for the bootstrap estimator smoothed in this way are summarised in the LSCV column of Table 1.

4.2 The BE Method

By comparison with the LSCV method, the BE (bootstrap estimation) method, as well as the methods in the following paragraph, is more directly focussed on the bootstrap problem in that it is based on the idea of minimising an estimate of the mean squared error (MSE) of the bootstrap estimator.

Ideally, if the true \(F\) were known, the best choice of \(h\) would be the one, \(h_{MSE}\) say, for which

\[
MSE_{\alpha}(F) = E_{\tilde{r}}[\alpha(\hat{F}_h) - \alpha(F)]^2
\]
is minimum. In practice, $F$ is unknown. The key idea here is that an estimate, $h_{BE}$ say, of $h_{MSE}$ can be obtained by minimising the bootstrap estimate

$$BE(h, g) = E_{	ilde{F}_g} \{ \alpha(\tilde{F}_g) - \alpha(\tilde{F}_g) \}^2$$

(7)

of $MSE\{\alpha(\tilde{F}_h)\}$. Here $\tilde{F}_g$, with $g$ to be specified, and $\tilde{F}_h$ are constructed by applying the kernel method to the sample data and to a random sample $Y_1, Y_2, \ldots, Y_n$ drawn from $\tilde{F}_g$ respectively. For $g = 0$, (7) becomes the standard bootstrap estimate of $MSE\{\alpha(\tilde{F}_0)\}$. Properties of such estimation methods for linear functionals are discussed in De Angelis & Young (1990), where it is argued that for the problem under consideration here it is appropriate to take $g = 0$.

Computationally, implementation of this method is much more expensive than LSCV. The function $BE(h; 0)$ cannot be computed directly and its estimation over a grid of 20 equally spaced values of $h$ in $[0, 1]$ was undertaken using a two-level bootstrap. A first level, consisting of the sampling of 50 datasets from $F_n$, was used to replace the expectation in (7) by a finite average. From the distribution $\tilde{F}_h$ determined by each of these 50 samples, a second level of 50 resamplings was drawn to estimate $\alpha(\tilde{F}_h)$.

A simple search amongst the 20 estimated values of $MSE$ gave the smoothing parameter $h$ for the bootstrap estimation. The estimated mean squared errors over the 100 simulations are displayed in the BE column of Table 1.

4.3 The PLUGIN and QUICKSTRAP Methods

The basic idea is the same as before: estimate $MSE\{\alpha(\tilde{F}_h)\}$ and use in the construction of the smoothed bootstrap estimator the value of $h$ that minimises such an estimate.

The so called PLUGIN technique studied here is in fact that presented by Young (1988). The only difference is in the definition (1), which is now simpler than the one used in Young (1988). Such a choice is motivated by the scale invariance of the correlation coefficient. Following Young (1988), the starting point is the observation that

$$\alpha(F) = (\text{var}_F Z)^t = n^{-1} \beta(F) + O(n^{-3/2})$$

(8)

where

$$\beta(F) = \left[ \frac{\rho^2}{(1 - \rho^2)^2} \left\{ \frac{\mu_{22}}{\mu_{11}^2} + \frac{1}{4} \left( \frac{\mu_{40}}{\mu_{20}} + \frac{\mu_{04}}{\mu_{02}} + \frac{2\mu_{22}}{\mu_{20}\mu_{02}} \right) \right\} + \left( \frac{\mu_{31}}{\mu_{11}^2 \mu_{20}} + \frac{\mu_{13}}{\mu_{11} \mu_{02}} \right) \right]^t$$

(9)

with $\mu_{ij} = \int x_i x_j dF(x_1, x_2)$, $\rho = \mu_{11} / \{\mu_{20}\mu_{02}\}^{1/2}$: see Kendall & Stuart (1977, p. 251). In order to simplify the notation we suppose here that $F$ has mean zero.
From (8) it is clear that estimation of \((\text{var} F Z)^t\) is approximately equivalent to that of \(\beta(F)\). Furthermore, by virtue of (3), the sampling properties of the smoothed bootstrap estimator \(\beta(\hat{F}_h)\) of \(\beta(F)\) are, to \(O_p(n^{-1})\), the same as those of the estimator

\[
A(\hat{F}_h) = n^{-1} \sum_{i=1}^{n} w^*(x_i)
\]

of a certain linear functional \(A(F) = \int a(t) dF(t)\), whose form can easily be derived from (9). In (10),

\[
w^*(x) = \int a(x + hV^{1/2}\xi)K(\xi)d\xi.
\]

Making the transformation \(Y = S^{-1}X\) where

\[
S = \begin{bmatrix} \mu_2 & 0 \\ 0 & \mu_0 \end{bmatrix} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix},
\]

the mean squared error of \(A(\hat{F}_h)\) in terms of the new variable \(Y\) is

\[
\text{MSE}[A(\hat{F}_h)] = [E_{F^*}[w^{**}(Y) - A(F)]]^2 + n^{-1}\text{var}_{F^*}[w^{**}(Y)]
\]

where \(F^*\) denotes the distribution of \(Y\) and

\[
w^{**}(Y) = \int a_5(Y + hS^{-1}V^{1/2}\xi)K(\xi)d\xi,
\]

with \(a_5(u) = a(Su)\). For a Gaussian kernel, it is possible to obtain a simple explicit form for (11) as

\[
w^{**}(Y) = \beta_0(Y_1/Y_2^2 + h^2Y_1^2 + h^2Y_2^2 + h^4 - m_{22}(Y_1^2 + Y_2^2) + \frac{1}{2}\rho(m_{11} - m_{13})(Y_1 - Y_2) - 2m_{22}h^2 - (m_{13} + m_{31})Y_1Y_2)
\]

with \(\beta_0 = \{2\beta(F)\}^{-1}\) and \(m_q = E_{F^*}(Y^q)\).

With use of computer algebra it is then easy to show that

\[
\text{MSE}[A(\hat{F}_h)] \propto n^{-1}(\gamma_0^2 + 4h^2\gamma_1 + 4h^2\gamma_2 - 16nh^6\gamma_3 + 4nh^8)
\]

(12)
where
\[
\gamma_0 = \rho^2(m_{13} - m_{31})^2(m_{04} + m_{40} - 2m_{22}) \\
+ 4\rho(m_{22}(m_{04} - m_{40})(m_{31} - m_{13}) + (m_{42} - m_{24}) \\
\times (m_{31} - m_{13}) + (m_{13}m_{31}^2 + m_{31}m_{13}^2 - m_{13}^3 - m_{31}^3)) \\
+ 4m_{22}^2(m_{04} + m_{40} + 2m_{22} - 1) + 12m_{22}(m_{13} + m_{31})^2 \\
- 8m_{33}(m_{13} + m_{31}) - 8m_{22}(m_{24} + m_{42}) + 4m_{44},
\]

\[
\gamma_1 = \rho(m_{04} - m_{40})(m_{13} - m_{31}) - 2m_{22}(m_{04} + m_{40}) \\
- 2(m_{13} + m_{31})^2 - 4m_{22}(m_{22} - 1) + 2(m_{24} + m_{42})
\]

\[
\gamma_2 = m_{04} + m_{40} + 4n(m_{22} - 1)^2 + 2m_{22} - 4
\]

\[
\gamma_3 = m_{22} - 1.
\]

The method proceeds now exactly as in Young (1988). The \(\gamma_i, i = 0, 1, 2, 3\) and then the expression (12) are estimated from the observed sample and the \(h\) minimising the estimated mean squared error is used in the bootstrap estimation. Once again the minimisation was performed using the NAG subroutine library.

More straightforward is the QUICKSTRAP procedure which derives from observing that from the above,

\[
MSE\{A(\hat{F}_h)\} \propto \gamma_0/n + (4\gamma_1/n)h^2 + (4\gamma_2^2)h^4 + O(h^6 + h^4/n)
\]  
(13)

with \(\gamma_0\) and \(\gamma_1\) as previously defined and \(\gamma_2^2 = 4(m_{22} - 1)^2\). The strategy works by estimating the leading terms in (13)

\[
MSE\{A(\hat{F})\} \propto \gamma_0/n + (4\gamma_1/n)h^2 + (4\gamma_2^2)h^4
\]

and minimising this estimate. Numerical minimisation of the estimated form of (14) is trivial compared with that of the estimated form of the full mean squared error: if \(\hat{\gamma}_1 > 0\) the minimising \(h\) is \(\hat{h} = 0\), otherwise \(h = \{-\hat{\gamma}_1/(2n\gamma_2^2)\}^{1/2}\).

Results corresponding to the two methods discussed here are reported in the columns PLUG and QUICK of Table 1 respectively. UNS refers to the unsmoothed bootstrap estimator of \(\alpha(F)\).

5. DISCUSSION

It is clear from the results that empirical smoothing of the bootstrap can lead to considerable reduction in the mean squared error of the bootstrap estimator, par-
particularly in smaller sample sizes and if the underlying distribution is Gaussian or nearly Gaussian. For the cases of a lognormal distribution or the Student's t distribution with 3 degrees of freedom smoothing is generally less effective.

Of the smoothing methods studied, the automatic and easily implemented LSCV method displays the best overall performance, even though it is not oriented towards the estimation problem under consideration. The LSCV method should not be expected to be as effective in general as it is in this problem: see De Angelis and Young (1990).

The other automatic smoothing method considered, the BE method, performs less effectively in this problem, leading to estimation which is uniformly poorer than that obtained from LSCV. The BE method may, of course, have been penalised somewhat in our study by the reduced number of simulations and resamplings used in its implementation: these were cut down deliberately to avoid otherwise enormous computational costs.

The results corresponding to the PLUGIN and QUICKSTRAP procedures are especially worth noting. These methods, which have considerable computational advantage over the BE procedure, which in a sense they approximate, are seen as more effective than the BE method, at least for smaller sample sizes. The performances of the PLUGIN and QUICKSTRAP procedures are particularly impressive when we bear in mind that they rely heavily on the MSE approximation (12). When the underlying distribution is $t_3$ not all of the moments required for (12) exist, while for the lognormal distribution, though the moments exist, they are very difficult to estimate accurately from small samples.

Our results here on the effectiveness of the PLUGIN and QUICKSTRAP methods bear on a comment made by DiCiccio and Romano (1988), who expressed doubt that the appropriate amount of smoothing for non-linear functionals might be successfully obtained from their linear approximations. Though they were concerned primarily with whether smoothing is effective in reducing coverage error in confidence intervals for the correlation, we have established here that examination of the linear approximation is a useful and simple approach to choosing the appropriate amount of smoothing when estimating sampling properties of the correlation coefficient.

References


