1 The Bootstrap

1.1 The Bootstrap Concept

To start with we shall just consider samples $Y = (Y_1, Y_2, ..., Y_n)$ and statistics $s(\mathbf{Y})$ calculated from them. The Y_i may themselves be vectors, and they are not necessarily independent, though often they will be. We shall assume initially however that the Y_i are mutually independent so that Y is a random sample. (The case of several samples is considered in Section () and the case when the Y_i are dependent will be discussed in Section ().)

Bootstrapping is simply a numerical way of answering the question: 'What is $G(.)$, the probability distribution of a statistic $s(Y)$, calculated from a sample $Y = (Y_1, Y_2, ..., Y_n)$?'

If we knew $F(.)$, the distribution of Y, the above question, 'What is $G(.)$?', is easily answered numerically. We simply generate a number, B, of independent samples of Y: $\mathbf{y}_1$, $\mathbf{y}_2$, ... $\mathbf{y}_B$, and calculate the

statistic $s_j = s(\mathbf{y}_j)$ from each sample. Then the *empirical distribution function* (EDF) of the sample $\mathbf{s} = (s_1, s_2, ..., s_B)$,

$$G_n(s \mid \mathbf{s}) = \frac{\#\text{ of } s_j \leq s}{B},$$

will converge pointwise with probability one to $G(s)$. [Here and in what follows we shall use the notation $G_n(.)$ to denote the EDF of a sample of size n drawn from the distribution $G(.)$. Where the precise sample $\mathbf{s}$ used to construct G_n needs to be made explicit then we use the notation $G_n(.|\ \mathbf{s})$ or $G_n(s|\ \mathbf{s})$] The basic sampling process is thus:

The Basic Sampling Process

For $j = 1$ to B
_For $i = 1$ to n
__Draw y_{ij} from $F(.)$
_Next i
_Calculate $s_j = s(\mathbf{y}_j)$
Next j
Form $G_n(.|\ \mathbf{s})$

The problem is that we do not know $F(.)$ usually. However we do have the EDF, $F_n(.|\ \mathbf{y})$, formed from the sample $\mathbf{y}$. The bootstrap idea is to use $F_n(.|\ \mathbf{y})$ instead of $F(.)$ in the above basic process. To obtain a sample we draw values, not from $F(.)$, but from $F_n(.|\ \mathbf{y})$. This is equivalent to drawing a sample of the same size n, *with replacement,* from the original set of y's. We call such a sample a *bootstrap sample* to distinguish it from the original sample, and write it as $\mathbf{y}^* = (y_1^*,\ y_2^*,\ ..., y_n^*)$. As in the basic process, B such bootstrap samples are drawn, and the statistic is calculated from these each of these samples:

$s_j^* = s(\mathbf{y}_j^*)$. The EDF, $G_n(. \mid \mathbf{s}^*)$, of these bootstrap statistics s_j^*, is our estimate of $G(.)$. The bootstrap process is as follows

The Bootstrap Sampling Process

Given a random sample $\mathbf{y} = (y_1, y_2, ...y_n)$ from $F(.)$

Form the EDF $F_n(.|\mathbf{y})$

For $j = 1$ to B

_For $i = 1$ to n

__Draw y_{ij}^* from $F_n(.|\mathbf{y})$

_Next i

_Calculate $s_j^* = s(\mathbf{y}_j^*)$

Next j

Form $G_n(.|\mathbf{s}^*)$

The underlying hope and expectation is that the bootstrap process will under many conditions reproduce the behaviour of the original process.

1.2 Basic Method

We shall assume that the objective is to use the random sample $\mathbf{y} = \{y_1, y_2, ..., y_n\}$ to estimate a parameter $\eta = \eta(F(.))$ that describes some quantity of interest of the distribution of Y. Typical examples are the mean

$$\eta = \int_{-\infty}^{\infty} y dF(y)$$

or some other moment of Y, or a quantile $\eta = q(p)$, defined by

$$\int_{-\infty}^{q(p)} dF(y) = p.$$

The statistic $s(Y)$ can then be regarded as any appropriate quantity that we may care to use for estimating η. We shall emphasise this view by using the alternative but entirely equivalent notation $\hat{\eta}(Y) \equiv s(Y)$, with the 'hat' indicating an extimated quantity. Thus, when η is the mean, an obvious statistic is the sample mean

$$s(\mathbf{Y}) = \bar{Y}.$$

It is important to realise that we do not have to use the sample version as the estimator, assuming we can calculate it at all. In the present example we might use a trimmed sample mean, or even the median rather than the sample mean. What we *shall* do is regard the sample version of the parameter as being the quantity of interest in the bootstrap process, i.e. the parameter of interest in the bootstrap process is

$$\eta^* = \eta(F_n(.))$$

Thus we think of the bootstrap process as one where we are generating bootstrap samples $\mathbf{y}_j^*$ from which bootstrap estimates $s_j^* = s(\mathbf{y}_j^*)$ are obtained that estimate the bootstrap parameter η^*. With this viewpoint, the bootstrap process is a prime example of the so-called *plug-in approach*, being a precise analogue of the original process with the only difference being that the known $F_n(.)$ is plugged-in for, i.e. replaces, the unknown $F(.)$. The *bootstrap principle* is that the bootstrap process reproduces the properties of the original basic process. Useful quantities that are

obtained from the bootstrap process are the sample mean and variance of the bootstrap estimates:

$$\bar{s}^* = \frac{1}{B}\sum_{j=1}^{B} s_j^*,$$

$$\sigma^{*2} = \frac{1}{B-1}\sum_{j=1}^{B}(s_j^* - \bar{s}^*)^2.$$

We also have an estimate of the bias of the mean of the bootstrap estimates: $b^* = \bar{s}^* - \eta^*$. A *bootstrap bias adjusted estimate* for η is thus $\breve{\eta} = \hat{\eta} - b^*$.If the bootstrap expectation satisfies $E^*(s_j^*) = \hat{\eta}$ (here E^* denotes the conditional expectation $E[s_j^*|\ F_n(.|\mathbf{y})]$, where $F_n(.)$ is treated as being the parent distribution), then $\bar{s}^* \to \hat{\eta}$ in probability as $B \to \infty$, and we have

$$\breve{\eta} = \hat{\eta} - b^* = \hat{\eta} - (\bar{s}^* - \eta^*) \to \hat{\eta} - (\hat{\eta} - \eta^*) = \eta^*.$$

Thus, in this case, adjusting the original estimator for bias using the bootstrap bias is equivalent to using the bootstrap parameter η^* as the initial estimate

1.3 The Double Bootstrap & Bias Correction

The bootstrap bias adjustment might not remove all the bias. At the expense of significantly increased computation, a second application of the bootstrap can be made. This is known as *the double bootstrap*, and consists of an outer bootstrap that is exactly the structure of the initial bootstrap, but where each step within this outer bootstrap is itself a (inner) bootstrap. The precise calculation is as follows.

The Double Bootstrap (for Bias Correction)

Given a random sample $\mathbf{y} = (y_1, y_2, ...y_n)$ from $F(.)$

Outer Bootstrap:

For $j = 1$ to B

_For $i = 1$ to n

__Draw y^*_{ij} from $F_n(.|\mathbf{y})$

_Next i

_Let $\eta^{**}_j = \eta(F_n(.|\mathbf{y}^*_j))$

_**Inner Bootstrap:**

_For $k = 1$ to B

__For $i = 1$ to n

___Draw y_{ijk}^{**} from $F_n(.|\mathbf{y}_j^*)$

__Next i

__Calculate $s_{jk}^{**} = s(\mathbf{y}_{jk}^{**})$

_Next k

_Let $\bar{s}_{j\cdot}^{**} = \frac{1}{B}\sum_{k=1}^{B} s_{jk}^{**}$

_Let $b_j^{**} = \bar{s}_{j\cdot}^{**} - \eta_j^{**}$

_Let $\breve{\eta}_j^* = \hat{\eta}_j^* - b_j^{**}$

_End of Inner Bootstrap

Next j

Let $\bar{\breve{\eta}}_{\cdot}^* = \frac{1}{B}\sum_{j=1}^{B} \breve{\eta}_j^*$

Let $b^* = \bar{\breve{\eta}}_{\cdot}^* - \eta^*$

Let $\breve{\eta} = \hat{\eta} - b^*$

End of Outer Bootstrap

Whilst this can lead to a reduction in the bias, the effect on the variance of the estimator is not so clear. The main obvious penalty is the increase in computing cost. If a full numerical method is used to carry

out the double bootstrap then the computation increases quadratically with B. The full double bootstrap is not always necessary. It may be possible to avoid the inner bootstrap by using an estimator of η that is already bias adjusted. For example if it is thought that the original estimator might seriously biased then one might replace it immediately with η^* and then use bootstrap bias adjustment on η^* using the single bootstrap. The single bootstrap yields an estimate of the variance of $\hat{\eta}^*$. When the bias correction is small this same variance will of course estimate the variance of $\breve{\eta}$ as well. However when the bias correction is large then the double bootstrap has to be used

1.4 Parametric Bootstrap

We now consider parametric bootstrapping. Again we focus on a random sample $\mathbf{y} = (y_1, y_2, ..., y_n)$ drawn from a distribution $F(y, \boldsymbol{\theta})$, with the difference that the form of F is known but which depends on a vector $\boldsymbol{\theta}$ of parameters that is unknown. We denote the unknown true value of $\boldsymbol{\theta}$ by $\boldsymbol{\theta}_0$. It is then natural to carry out bootstrap sampling, not from the EDF $F_n(y \mid \mathbf{y})$, but from the distribution $F(y, \hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ is an estimate of $\boldsymbol{\theta}_0$. An obvious choice for $\hat{\boldsymbol{\theta}}$ is the *maximum likelihood* (ML) estimator because it has an asymptotic distribution that is easy to calculate. We define the log-likelihood by

$$L(\boldsymbol{\theta}, \mathbf{y}) = \sum_{i=1}^{n} \log f(y_i; \boldsymbol{\theta}). \tag{1}$$

where $\mathbf{y} = (y_1, y_2, ..., y_n)$. The method of ML is to estimate $\boldsymbol{\theta}_0$ by maximizing L, treated as a function of $\boldsymbol{\theta}$, subject to $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is some allowed region of the parameter space. for the remainder of the subsection $\hat{\boldsymbol{\theta}}$ will represent the ML estimator. Regularity conditions which ensure that standard asymptotic

theory applies are given in Wilks (1962) or Cox and Hinkley (1974) for example. When they apply $\hat{\boldsymbol{\theta}}$ can found by solving $\left.\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\hat{\boldsymbol{\theta}}} = \mathbf{0}$. Moreover

$$-n^{-1} \left.\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\right|_{\boldsymbol{\theta}_0} \rightarrow I(\boldsymbol{\theta}_0), \tag{2}$$

a positive definite matrix, and the asymptotic distribution of $\hat{\boldsymbol{\theta}}$, as $n \rightarrow \infty$, is

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \text{ is } N(\mathbf{0}, V(\boldsymbol{\theta}_0)) \tag{3}$$

where $V(\boldsymbol{\theta}_0) = [I(\boldsymbol{\theta}_0)]^{-1}$. From (2) a reasonable approximation for $I(\boldsymbol{\theta}_0)$ is

$$I(\hat{\boldsymbol{\theta}}) \simeq -n^{-1} \left.\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\right|_{\hat{\boldsymbol{\theta}}}. \tag{4}$$

More generally suppose that η is a smooth function of $\boldsymbol{\theta}$, ie $\eta = \eta(\boldsymbol{\theta})$. The ML estimate of η is then simply

$$\hat{\eta} = \eta(\hat{\boldsymbol{\theta}}),$$

where $\hat{\boldsymbol{\theta}}$ is the ML estimate of $\boldsymbol{\theta}$. The asymptotic normality of $\hat{\eta}$ follows from that of $\hat{\boldsymbol{\theta}}$, assuming that

$\eta(\boldsymbol{\theta})$ can be approximated by a linear Taylor series in $\boldsymbol{\theta}$. This method of obtaining the asymptotic behaviour of $\hat{\eta}$ is called the *delta method*. We have

$$n^{1/2}(\hat{\eta} - \eta_0) \sim N(0, \sigma^2(\boldsymbol{\theta}_0)) \tag{5}$$

with

$$\sigma^2(\boldsymbol{\theta}_0) = \mathbf{g}^T(\boldsymbol{\theta}_0) V(\boldsymbol{\theta}_0) \mathbf{g}(\boldsymbol{\theta}_0) \tag{6}$$

where

$$\mathbf{g}(\boldsymbol{\theta}_0) = \left. \frac{\partial \eta(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}_0}. \tag{7}$$

An approximation for the distribution of $\hat{\eta}$ is given by (5) with $\hat{\boldsymbol{\theta}}$ used in place of the unknown $\boldsymbol{\theta}_0$. The above asymptotic theory is firmly established and much used in practice. For this reason parametric bootstrapping is not especially needed when parametric models are fitted to large samples. However, for small samples Monte-Carlo simulation is a well known method for examining the behaviour of estimators and fitted distributions in parametric problems . Such methods are actually parametric bootstrapping under a different name and as such have been known and used for

a long time. In this sense parametric bootstrapping significantly predates the direct non-parametric bootstrap. Parametric bootstrapping lacks the computational novelty of direct sampling in that resampling is not involved. Perhaps it is because of this that the significance of parametric bootstrapping was not fully appreciated, and its theoretical importance not recognized until the landmark paper of Efron (1979). If we view bootstrapping as a numerical sampling approach involving models that may include parameters, then its potential is significantly extended.

2 Percentiles and Confidence Intervals

2.1 Percentiles

We shall denote the estimate of a percentile q_p, with percentile probability p, by $\hat{q}_p(\mathbf{y})$. The obvious choice for estimating a percentile non-parametrically is to use the corresponding percentile of the EDF. There is a certain ambiguity in doing this as the EDF is a step function. Thus all percentiles in the interval $(\frac{k-1}{n}, \frac{k}{n})$ are estimated by the observed order statistic $y_{(k)}$. Conversely the percentile $\frac{k}{n}$ is estimated by any point y for which $F_n(y) = k/n$. Such ambiguities are removed if the EDF is smoothed. One way is to use a kernel estimator of the density but in most simulations this is perhaps unnecessarily elaborate. The simplest smoothing procedure is to note that if $Y_{(k)}$ is the kth order statistic (as opposed to the observed value $y_{(k)}$) then $E[F(Y_{(k)})] = k/(n+1)$ and to use this value to estimate the value of F at the observed order statistic $y_{(k)}$. We can now simply interpolate between these estimated points $(y_{(k)},\ k/(n+1))$ for $k = 1, 2, ..., n$.

The range can be extended to $(0, 0)$ (using the line segment joining $(0, 0)$ and $(y_{(1)},\ 1/(n+1))$ if Y is known to be a positive random variable. If we denote this smoothed version of $F_n(.)$ by $\tilde{F}_n(.)$ then an obvious estimator for $\hat{q}_p(\mathbf{y})$ is

$$\hat{q}_p(\mathbf{y}) = \tilde{F}_n^{-1}(p)$$

Estimating $F(.)$ in the range $(y_{(n)}, \infty)$, or equivalently q_p for $p > n/(n+1)$ is not advisable unless the tail behaviour of F is known. The bootstrap analogue of $\hat{q}_p(\mathbf{y})$ is obtained in the usual way. We draw $\mathbf{y}^*$, a bootstrap sample from $\mathbf{y}$, and construct $\tilde{F}_n(.|\ \mathbf{y}^*)$, the smoothed version of the bootstrap EDF. [Where there is no ambiguity we shall write $F_n^*(.)$ for $F_n(.|\ \mathbf{y}^*)$ and $\tilde{F}_n^*(.)$ for $\tilde{F}_n(.|\ \mathbf{y}^*)$.] We can now calculate $\hat{q}_p^*$ as $\hat{q}_p^* = \tilde{F}_n^{*-1}(p)$. The bootstrap parameter here is $\eta^* = \hat{q}_p(\mathbf{y})$. In the parametric case, things are much easier. Quantiles are simply estimated from the fitted distribution $F(., \hat{\boldsymbol{\theta}})$

2.2 Confidence Intervals by Direct Bootstrapping

We now consider the construction of confidence intervals for a parameter of interest, η, and consider to what extent the corresponding bootstrap process. can be used to supply a confidence interval for this original parameter η. The key requirement is that the distribution of the bootstrap difference $\hat{\eta}^* - \eta^*$ should be close to that of $\hat{\eta} - \eta$. Roughly speaking we need

$$P^*(\sqrt{n}(\hat{\eta}^* - \eta^*) \leq y) - P(\sqrt{n}(\hat{\eta} - \eta) \leq y) \rightarrow 0 \quad (8)$$

for any y as $n \rightarrow \infty$. (As before with E^*, P^* denotes the conditional probability evaluation treating $F_n(.|\,\mathbf{y})$ as the parent distribution.) More precise statements of this convergence (8) and other results are given in Section (). We can proceed as follows. We generate B bootstrap samples $\mathbf{y}_j^*$, $j = 1, 2, ..., B$ and corresponding bootstrap estimates $\hat{\eta}_j^*$, $j = 1, 2, ..., B$. Then we select an appropriate confidence level $(1-2\alpha)$ (we use 2α rather than α so that α corresponds to each of the

two tail probabilities in two-sided intervals) and find values a^* and b^* for which

$$P^*(a^* \leq \hat{\eta}^* - \eta^* \leq b^*) \simeq 1 - 2\alpha.$$

We then appeal to (8) to replace $\hat{\eta}^* - \eta^*$ by $\hat{\eta} - \eta$ and obtain

$$P(a^* \leq \hat{\eta} - \eta \leq b^*) \simeq 1 - 2\alpha$$

which on inversion gives the approximate $(1 - 2\alpha)$ confidence interval

$$\hat{\eta} - b^* \leq \eta \leq \hat{\eta} - a^* \tag{9}$$

for η. One way of obtaining a^* and b^* is as follows. Let the bootstrap estimates of the α and $(1 - \alpha)$ quantiles obtained from the smoothed EDF of the $\hat{\eta}_j^*$'s be $\hat{q}_\alpha(\hat{\boldsymbol{\eta}}^*)$ and $\hat{q}_{1-\alpha}(\hat{\boldsymbol{\eta}}^*)$. These are estimates of the quantiles of the distribution of $\hat{\eta}^*$; so the corresponding quantiles of $\hat{\eta}^* - \eta^*$ are

$$a^* = \hat{q}_\alpha(\hat{\boldsymbol{\eta}}^*) - \eta^* \text{ and } b^* = \hat{q}_{1-\alpha}(\hat{\boldsymbol{\eta}}^*) - \eta^*. \tag{10}$$

Substituting these into (9) and noting that $\eta^* = \hat{\eta}$ gives what is normally called *the basic bootstrap confidence interval.*

$$2\hat{\eta} - \hat{q}_{1-\alpha}(\hat{\boldsymbol{\eta}}^*) \leq \eta \leq 2\hat{\eta} - \hat{q}_{\alpha}(\hat{\boldsymbol{\eta}}^*). \tag{11a}$$

This confidence interval is also known as the *hybrid bootstrap confidence interval* as we have essentially replaced percentage points of the unknown original EDF by those of the smoothed bootstrap EDF. There is substantial evidence that this basic bootstrap can be significantly improved (in terms of giving coverage that is closer to the nominal level) if we use a *pivotal* quantity, like a studentized mean, rather than focusing just on the difference $\hat{\eta} - \eta$. The reason is that a pivotal quantity is less dependent (ideally not dependent at all)) on the form of the original distribution. This would mean that the confidence interval is less influenced by any difference there may be in the distributions of $\hat{\eta}^* - \eta^*$ and $\hat{\eta} - \eta$. We consider this next.

2.3 Studentization

Suppose that $\hat{\sigma}^2(\mathbf{y})$, an estimate of the variance of $\hat{\eta}$, can be calculated from the sample $\mathbf{y}$. This will be the case for $\hat{\eta} = \bar{y}$ say, when we have $\hat{\sigma}^2(\mathbf{y}) = s^2$, the sample variance. As before suppose we can find a^* and b^* such that

$$P^*(a^* \leq (\hat{\eta}^* - \eta^*)/\hat{\sigma}^* \leq b^*) \simeq 1 - 2\alpha.$$

If we then appeal to a similar result to (8) and replace $(\hat{\eta}^* - \eta^*)/\hat{\sigma}^*$ by $(\hat{\eta} - \eta)/\hat{\sigma}$, we obtain

$$P(\hat{\sigma}a^* \leq \hat{\eta} - \eta \leq \hat{\sigma}b^*) \simeq 1 - 2\alpha$$

which on inversion gives the approximate $(1 - 2\alpha)$ confidence interval

$$\hat{\eta} - \hat{\sigma}b^* \leq \eta \leq \hat{\eta} - \hat{\sigma}a^*.$$

We can now calculate a^* and b^* by drawing B bootstrap versions of $z = (\hat{\eta} - \eta)/\hat{\sigma} : z_j^* = (\hat{\eta}_j^* - \eta^*)/\hat{\sigma}_j^*$, $j = 1, 2, ..., B$. Let $\hat{q}_\alpha(\mathbf{z}^*)$ and $\hat{q}_{1-\alpha}(\mathbf{z}^*)$ be the quantiles obtained from the EDF of these z_j^*. The confidence interval for η is now

$$(\hat{\eta}_\alpha^{Stud},\ \hat{\eta}_{1-\alpha}^{Stud}) = (\hat{\eta} - \hat{\sigma}\hat{q}_{1-\alpha}(\mathbf{z}^*),\ \hat{\eta} - \hat{\sigma}\hat{q}_\alpha(\mathbf{z}^*)). \tag{12}$$

This is usually called the *studentized bootstrap confidence interval*. Studentized bootstrap intervals can be readily used with the parametric model $F(y, \boldsymbol{\theta})$ of Section (). Suppose that $\mathbf{y}$ is a random sample of size n drawn from the distribution $F(y, \boldsymbol{\theta}_0)$, and that $\eta_0 = \eta(\boldsymbol{\theta}_0)$ is the quantity of interest. We can estimate η_0 using $\hat{\eta} = \eta(\hat{\boldsymbol{\theta}})$ where $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}_0$.When n is large we can use the asymptotic approximation

$$n^{1/2}(\hat{\eta} - \eta_0)/\sigma(\hat{\boldsymbol{\theta}}) \sim N(0, 1).$$

When n is not large it is worth employing bootstrapping. We draw B bootstrap samples $\mathbf{y}_j^*$ $j = 1, 2, ..., B$ from the fitted distribution $F(y, \hat{\boldsymbol{\theta}})$. From each sample we obtain the bootstrap ML estimator $\hat{\boldsymbol{\theta}}_j^*$ and bootstrap studentized quantity $z_j^* = n^{1/2}(\hat{\eta}_j^* - \eta^*)/\sigma(\hat{\boldsymbol{\theta}}_j^*)$. The quantiles $\hat{q}_\alpha(\mathbf{z}^*)$ and $\hat{q}_{1-\alpha}(\mathbf{z}^*)$ can be obtained from the EDF of these z_j^* and the studentized interval (12) constructed. In the non-parametric case an estimate of the variance of $\hat{\eta}$ may not be immediately available. One possibility is to use the double bootstrap as given previously (where $s = \hat{\eta}$) with the inner

loop taking the form

_**Inner Bootstrap:**

_For $k = 1$ to B

__For $i = 1$ to n

___Draw y_{ijk}^{**} from $F_n(.|\mathbf{y}_j^*)$

__Next i

__Let $s_{jk}^{**} = s(\mathbf{y}_{jk}^{**})$

_Next k

_Let $\hat{\eta}_j^* = s(\mathbf{y}_j^*)$

_Let $\overline{s}_{j\cdot}^{**} = \frac{1}{B}\sum_{k=1}^{B} s_{jk}^{**}$

_Let $v_j^{**} = \frac{1}{B-1}\sum_{k=1}^{B}(s_{jk}^{**} - \overline{s}_{j\cdot}^{**})^2$

_Let $z_j^* = n^{1/2}(\hat{\eta}_j^* - \eta^*)/(v_j^{**})^{1/2}$

_**End of Inner Loop**

If the double bootstrap is considered computationally too expensive, then an alternative using *influence functions* can be used provided our statistic is expressible as functional of the EDF, i.e. $s(\mathbf{y}) = \eta[F_n(.|\mathbf{y})]$. Such statistics are termed *statistical functions*, and

were introduced by von Mises (1947). We then assume that the relevant functional $\eta = \eta[F(.)]$ has a linearised form (akin to a first order linear generalisation of a Taylor series): $\eta(G(.)) \simeq \eta(F(.)) + \int L_\eta(y|F(.))dG(y)$. Here

$$\begin{aligned} L_\eta(y|F(.)) &= \frac{\partial \eta\{(1-\varepsilon)F(.) + \varepsilon H_y(.)\}}{\partial \varepsilon} \\ &= \lim_{\varepsilon \to 0} \varepsilon^{-1}[\eta\{(1-\varepsilon)F(.) + \varepsilon H_y(.)\} \\ &\quad -\eta(F(.))], \end{aligned} \tag{13}$$

the derivative of η at F, is called the *influence function* of η, and $H_y(x)$ is the Heaviside unit step function with jump from 0 to 1 at $x = y$. The sample approximation

$$l(y|\mathbf{y}) = L_\eta(y|F_n(.|\mathbf{y})) \tag{14}$$

is called the *empirical influence function*. An analogous argument to that used in the delta method yields the non-parametric analogue of (6) as

$$\text{Var}[\eta\{F(.)\}] = n^{-1} \int L_\eta^2\{y|F(.)\}dF(y),$$

with sample version

$$\text{Var}[\eta\{F_n(.|\mathbf{y})\}] = n^{-2}\sum_{j=1}^{n} l^2(y_j|\mathbf{y}),$$

where the $l(y_j|\mathbf{y})$ are the *empirical function values* evaluated at the observed values y_j. In practice these values have to be evaluated numerically using (13) with, typically, $\varepsilon = (100n)^{-1}$. The ease with which the empirical influence function can be evaluated is rather problem dependent, and the potential complexity of this calculation detracts from this approach. One simple approximation is the jacknife approximation

$$l_{jack}(y_j|\mathbf{y}) = (n-1)(\hat{\eta}(\mathbf{y}) - \hat{\eta}(\mathbf{y}_{\setminus j}))$$

where $\hat{\eta}(\mathbf{y}_{\setminus j})$ is the estimate calculated from the sample $\mathbf{y}$ but with the observation y_j omitted. (See Davison and Hinkley 1997, problem 2.18.)

2.4 Percentile Methods

Suppose now that the distribution of $\hat{\eta}$ is symmetric, or more generally, there is some transformation $w = h(\hat{\eta})$ for which the distribution is symmetric. Examination of the form of the bootstrap percentiles (10), (see for example Hjorth, 1994, Section 6.6.2) shows that they do not depend on the explicit form of $h(.)$ at all. Instead we find that they can, with a change of sign, be swapped, ie a^* is replaced by $-b^*$ and b^* by $-a^*$. Then (9) becomes

$$(\hat{\eta}_{\alpha}^{Per},\ \hat{\eta}_{1-\alpha}^{Per}) = (\hat{q}_{\alpha}(\hat{\boldsymbol{\eta}}^*),\ \hat{q}_{1-\alpha}(\hat{\boldsymbol{\eta}}^*)). \qquad (15)$$

This $(1-2\alpha)$ confidence interval is known as the *bootstrap percentile interval*. This confidence interval is easy to implement but can be improved. We give two variants called the *bias corrected* (BC) and *accelerated bias corrected* (BCa) *percentile methods.* The basic idea in the BCa method is that there is a studentized transformation g for which

$$\frac{g(\hat{\eta}) - g(\eta)}{1 + ag(\eta)} + b \sim N(0,1). \qquad (16)$$

The BC method is simply the special case where $a = 0$. If we calculate a one sided confidence interval for η under the assumption (16) we find that if we set

$$\beta_1 = \Phi\left(b + \frac{b + z_\alpha}{1 - a(b + z_\alpha)}\right)$$

where $\Phi(.)$ is the standard normal distribution function, then with confidence $(1 - \alpha)$

$$\hat{q}_{\beta_1}(\hat{\boldsymbol{\eta}}^*) < \eta$$

where $\hat{q}_\beta(\hat{\boldsymbol{\eta}}^*)$ is the bootstrap percentile of the $\hat{\boldsymbol{\eta}}^*$ with probability β. In a similar way a two sided $1-2\alpha$ confidence interval is

$$\hat{q}_{\beta_1}(\hat{\boldsymbol{\eta}}^*) < \eta < \hat{q}_{\beta_2}(\hat{\boldsymbol{\eta}}^*)$$

where

$$\beta_2 = \Phi\left(b + \frac{b + z_{1-\alpha}}{1 - a(b + z_{1-\alpha})}\right).$$

The bias parameter b is obtained as

$$b = \Phi^{-1}\{\breve{F}_n(\hat{\eta} \mid \hat{\boldsymbol{\eta}}^*)\}$$

where $\tilde{F}_n(\hat{\eta} \mid \hat{\boldsymbol{\eta}}^*)$ is the smoothed EDF of the bootstrap sample $\hat{\boldsymbol{\eta}}^*$ evaluated at the original estimated value $\hat{\eta}$. If we now set $a = 0$ this gives the BC method. We denote the resulting two-sided version of the confidence interval by $(\hat{\eta}_\alpha^{BC}, \hat{\eta}_{1-\alpha}^{BC})$. The acceleration parameter a is arguably less easily calculated. Efron (1987) suggests the approximation (which is one-sixth the estimated standard skewness of the linear approximation to η) :

$$a = \frac{1}{6} \frac{\sum_{i=1}^n l^3(y_i|\mathbf{y})}{[\sum_{i=1}^n l^2(y_i|\mathbf{y})]^{3/2}}$$

where $l(y_i|\mathbf{y})$ are the values of the empirical influence function (14) evaluated at the observations y_i. We denote the two-sided version of the confidence interval given by this BCa method by $(\hat{\eta}_\alpha^{BCa}, \hat{\eta}_{1-\alpha}^{BCa})$.

3 Theory

Like many statistical methods, understanding of the practical usefulness of the bootstrap as well as its limitations has been built up gradually with experience through applications. This practical experience has been underpinned by a growing asymptotic theory which provides a basis for understanding when bootstrapping will work and when it will not. A truly general theory rapidly becomes very technical and is still incomplete. A detailed treatment is not attempted here, but some of the more accessible and useful results are summarized. Bootstrapping relies on sampling from the EDF $F_n(.|\mathbf{y})$ reproducing the behaviour of sampling from the original distribution $F(.)$. We therefore need convergence of $F_n(.|\mathbf{y})$ to $F(.)$. This is confirmed by the Glivenko-Cantelli theorem which guarantees strong convergence, i.e.

$$\sup_y |F_n(y|\mathbf{y}) - F(y)| \to 0 \text{ with probability } 1$$

as $n \to \infty$. Though reassuring this does not throw any direct light on the bootstrap process. To investigate

the bootstrap process we look instead at the equally well-known result that the process

$$Z_n(y) = \sqrt{n}(F_n(y) - F(y))$$

converges in probability to the Brownian bridge $B(F(y))$ as $n \to \infty$. [A Brownian bridge $B(t)$, $0 < t < 1$, is a Gaussian process for which $B(0) = B(1) = 0$, $E[B(u)] = 0$ and $E[B(u)B(v)] = u(1-v)$ for $0 < u < v < 1$.] Bickel and Freedman (1981) give the following bootstrap version of this.

Theorem Let $Y_1, Y_2, ...$ be a sequence of independent observations from $F(.)$, with $\mathbf{Y}_n$ comprising the first n values. Then for almost all such sequences (i.e. with probability 1) the bootstrap process

$$Z_n^*(y) = \sqrt{n}(F_n^*(y|\mathbf{Y}_n) - F_n(y))$$

converges in probability to the Brownian bridge $B(F(y))$ as $n \to \infty$. An interesting immediate application of this result to the calculation of confidence bands for $F(y)$ is given by Bickel and Freedman. We set a confidence level $1 - \alpha$ and, from the bootstrap process,

we then select $c_n(\alpha)$ so that, as $n \to \infty$,

$$P^*\{\sup_y |Z_n^*(y)| \leq c_n(\alpha)\} \to 1 - \alpha.$$

Then, as $Z_n^*(y)$ and $Z_n(y)$ converge to the same process $B(F(y))$, we have also that

$$P^*\{\sup_y |Z_n(y)| \leq c_n(\alpha)\} \to 1 - \alpha.$$

Thus an asymptotically correct $(1 - \alpha)$ confidence band for $F(y)$ is

$$F_n(y) \pm \frac{c_n(\alpha)}{\sqrt{n}}.$$

We now consider the distribution of an estimator $\hat{\eta}$ of a parameter η. We need the distributions of the bootstrap estimator $\hat{\eta}^*$ and of the difference $\hat{\eta}^* - \eta^*$ to reproduce respectively the distributions of $\hat{\eta}$ $(\equiv \eta^*)$ and of $\hat{\eta} - \eta$. The most general results obtained to date assume that the statistic can be expressed as $s = \eta(F_n)$ where both the EDF F_n and the underlying distribution F belong to a space $\mathcal{F}$ of distributions. Moreover such results require s to be differentiable in such a

way that a linear approximation can be formed with, as first order terms, the previously described influence functions, i.e.

$$\eta(F_n) = \eta(F) + n^{-1} \sum_{j=1}^{n} l(y_j|\mathbf{y}) + o_p(n^{-1/2}).$$

Then it is often possible to establish the analogous result for the bootstrap version

$$\eta(F_n^*) = \eta(F) + n^{-1} \sum_{j=1}^{n} l(y_j^*|\mathbf{y}^*) + o_p(n^{-1/2}),$$

and, provided Var$[\eta(F)]$ is finite, then, in probability, the sequence $y_1, y_2, ...$ is such that

$$\sup_s |P^*(\eta(F_n^*) \leq s) - P(\eta(F_n) \leq s)| \to 0$$

as $n \to \infty$. Here we have written $P^*(\eta(F_n^*) \leq s)$ for $P(\eta(F_n^*) \leq s \mid F_n)$. If, in addition, s is *continuously* differentiable, then we can usually strengthen the result to almost sure convergence. A detailed discussion is given by Shao and Tu (1995) who give examples of $\hat{\eta}(F)$ that satisfy such conditions. Convergence has

been studied in detail for the particular case of means and percentiles. In this case a more accessible approach using the Berry-Esséen inequality is possible, and this we discuss in the next section.

3.1 Convergence Rates

Efron (1979) considered the finite case where Y takes just a finite set of values which we can take to be $\{1, 2, ..., m\}$ with probabilities $\mathbf{p} = (p_1, p_2, ..., p_m)^T$. We shall not discuss this particular case, a good summary for which is provided by Hjorth (1994). We shall instead focus on the continuous case where $\eta = \mu \equiv E(Y)$, with $\hat{\eta} = \bar{Y}_n = \sum_{i=1}^{n} Y_i/n$. We also write $Var(Y) = \sigma^2$, $\rho = E(|Y|^3)$. The well-known Berry-Esséen theorem (see Serfling, 1980, for example) then provides a powerful result for investigating bootstrap convergence. To avoid long formulas we shall write:

$$H_n(y) = P(\sqrt{n}(\bar{Y}_n - \mu) \leq y)$$

for the (unknown) CDF of interest and

$$H_n^*(y) = P^*(\sqrt{n}(\bar{Y}_n^* - \bar{Y}_n) \leq y)$$

for the bootstrap version of this. We write

$$\tilde{H}_n(y) = P(\sqrt{n}\frac{(\bar{Y}_n - \mu)}{\sigma} \leq y)$$

for the (unknown) *standardized* CDF of interest,

$$\breve{H}_n(y) = P(\sqrt{n}\frac{(\bar{Y}_n - \mu)}{\hat{\sigma}_n} \leq y)$$

for the studentized CDF, and

$$\tilde{H}^*_n(y) = P^*(\sqrt{n}\frac{(\bar{Y}^*_n - \bar{Y}_n)}{\hat{\sigma}_n} \leq y)$$

for its bootstrap *studentized* version. We write also $\Phi(y)$ for the CDF of the standard normal distribution. The Berry-Esséen theorem states that if $Y_1, Y_2, ...$ are independent identically distributed random variables and $\rho = E[\left|Y_i^3\right|] < \infty$, then, for all n

$$\sup_y \left| H_n(y) - \Phi(\frac{y}{\sigma}) \right| < K\frac{\rho}{\sigma^3\sqrt{n}}.$$

The value of the constant given by Berry and Esséen, $K = \frac{33}{4}$, has been improved and reduced to .7975 (van Beeck,1972). Thus the theorem gives a bound on the effectiveness of the normal distribution in approximating $H_n(.)$. A typical application of the theorem is in proving the following.

Theorem If $\rho = E[\left|Y_i^3\right|] < \infty$ then, as $n \to \infty$,

$$\sup_y |H_n^*(y) - H_n(y)| \to 0 \tag{17}$$

for almost all sequences $y_1, y_2, ...$of independent observations drawn from $F(.)$.

Proof Outline. By the Berry-Esséen theorem we have

$$\left|P(H_n(y) - \Phi(\frac{y}{\sigma})\right| < K\frac{\rho}{\sigma^3\sqrt{n}}$$

and

$$\left|P^*(H_n^*(y) - \Phi(\frac{y}{\hat{\sigma}_n})\right| < K\frac{\rho_n^*}{\hat{\sigma}_n^3\sqrt{n}}.$$

By the strong law of large numbers $\mu^* \equiv \bar{y}_n \to \mu$, $\hat{\sigma}_n \to \sigma$, $\rho_n^* \to \rho$ all with probability 1. This shows that the two probabilities in (17) converge to the same normal distribution from which the result follows easily. □ In fact the above result does not depend on $\rho < \infty$; the condition $\sigma^2 < \infty$ is actually enough (see Bickel and Freedman, 1981 or Singh 1981.) However the condition $\rho < \infty$ and explicit use of the Berry-Esséen theorem is needed to establish the *rate*

of convergence more precisely. In the remainder of this section we summarize results obtained by Singh (1981) and Bickel and Freedman (1981). We need one further definition. A random variable has distribution $F(.)$ that is *lattice* if there is zero probability of it taking values outside the discrete points $y_j = c + jd$ $j = 0, 1, 2, ...$ where c and d are constants.

Theorem For almost all sequences Y_1, Y_2, ...of independent observations drawn from $F(.)$:

(i) If $E(Y^4) < \infty$, then

$$\limsup_{n\to\infty}\left[\frac{\sqrt{n}}{\sqrt{\log\log n}}\sup_y |H_n^*(y) - H_n(y)|\right]$$

$$= \frac{\sqrt{Var[(Y-\mu)^2]}}{2\sigma^2\sqrt{\pi e}}$$

(ii) If $E[|Y|^3] < \infty$ and $F(.)$ is lattice then

$$\limsup_{n\to\infty}\left[\sqrt{n}\sup_y \left|\tilde{H}_n^*(y) - \tilde{H}_n(y)\right|\right] = \frac{d}{\sqrt{2\pi}\sigma}$$

(iii) If $E[|Y|^3] < \infty$ and $F(.)$ is non-lattice then

$$\sqrt{n}\sup_y \left|\tilde{H}_n^*(y) - \tilde{H}_n(y)\right| \to 0$$

The result (i) shows that when $E(Y^4) < \infty$ then the convergence rate of H_n^* is $O(\sqrt{\log\log n/n})$. This is actually the same as that for approximating $\sqrt{n}(\bar{Y}_n - \mu)/\hat{\sigma}_n$, the *studentized* original $\bar{Y}$ by Z, a standard normal, i.e. approximating $H_n(y)$ by $\Phi(y/\hat{\sigma}_n)$. So in this case we do no better using the bootstrap than standard normal theory. Results (ii) and (iii) show that when $E[|Y|^3] < \infty$, then the convergence rate of $\tilde{H}_n^*(.)$ to $\tilde{H}_n(.)$ is at least $O(n^{-1/2})$ and is $o(n^{-1/2})$ when F is non-lattice. Thus bootstrapping does better than the usual normal approximation in this case. *This is the key theoretical result that underpins much of bootstrapping.* We consider now quantile estimation. Perhaps the most informative general result is that due to Singh (1981). Let $\eta = q(p) = F^{-1}(p)$; where the conditions of the theorem ensure that the inverse is uniquely defined. We can estimate η from the EDF using $\hat{\eta} = F_n^{-1}(p) = \sup\{y : F_n(y) \leq p\}$ and define $\hat{\eta}^* = F_n^{*-1}(p)$ from $\mathbf{y}^*$ as usual. We take $H_n = P(\sqrt{n}(\hat{\eta} - \eta) \leq y)$ and $H_n^* = P^*(\sqrt{n}(\hat{\eta}^* - \hat{\eta}) \leq y)$

Theorem If F has a bounded second derivative in a neighborhood of η and $f(\eta) = dF/dy|_\eta > 0$, then

$$\limsup_{n\to\infty} \frac{n^{1/4} \sup_y |H_n^*(y) - H_n(y)|}{\sqrt{\log\log n}} = c_{p,F} \text{ with prob. 1}$$

This result shows that $H_n^*(y)$ converges to $H_n(y)$ at rate $O(n^{-1/4} \log\log n)$. Now for a percentile estimator we have from the Berry-Esséen theorem

$$\sup_y \left| H_n(y) - \Phi(\frac{y}{\tau}) \right| = O(n^{-1/2})$$

where $\tau = \sqrt{p(1-p)}/f(\eta)$. Thus if we were to use a normal approximation for $H_n(y)$ we would use $\Phi(y/\hat{\tau})$ where $\hat{\tau}$ is an estimate of τ. Whether this is better than $H_n^*(y)$ thus depends on the converence rate of $\hat{\tau}$, and the matter is not a clear one.

3.2 Asymptotic Accuracy of EDF's

Edgeworth expansions are asymptotic series aimed at improving the normal approximation by introducing additional terms that try to correct for effects of skewness, kurtosis and higher moment which slow the rate of convergence to normality. For fixed n, asymptotic series usually diverge as more and more terms are included. However for a fixed number of terms, k say, the series converges as $n \to \infty$. The rate of convergence is usually of smaller order of than the last included term. We shall only consider the special case $\hat{\eta} = \bar{Y}_n$. Here, the general Edgeworth expansion is a power series in $n^{-1/2}$ and has the form:

$$\begin{aligned} \tilde{H}_n(y) &= \Phi(y) + n^{-1/2} p^{(1)}(y)\phi(y) + .. \\ &\quad + n^{-k/2} p^{(k)}(y)\phi(y) + o(n^{-k/2}) \end{aligned} \tag{18}$$

where $\phi(y) = (2\pi)^{-1/2} \exp(-y^2/2)$ is the standard normal density, and p_j is a polynomial of degree $3j-1$. We have explicitly

$$p^{(1)}(y) = -\frac{1}{6}\kappa_3(y^2 - 1)$$

and

$$p^{(2)}(y) = -\{\frac{1}{24}\kappa_4(y^2-3) + \frac{1}{72}\kappa_3^2(y^4-10y^2+15)\}$$

where κ_3 and κ_4 are the skewness and kurtosis of $F(.)$. The term involving $p^{(1)}$ corrects for the main effect of skewness, the term involving $p^{(2)}$ corrects for the main effect of kurtosis and for the secondary effect of skewness. Often the remainder term $o(n^{-k/2})$ can be replaced by $O(n^{-(k+1)/2})$ when the Edgeworth expansion is said to be $(k+1)$th order accurate. Usually inclusion of more than one or two correction terms becomes counter-productive as the coefficients associated with the powers of $n^{-1/2}$ rapidly become large with k. When $E(|Y|^3) < \infty$ and Y is non-lattice then one-term Edgeworth expansions for both $H_n(y)$ and $H_n^*(y)$ exist and a comparison (see Shao and Tu, 1995, Section 3.3.3) shows that $H_n^*(y)$ has smaller asymptotic mean square error than $\Phi(y/\hat{\sigma}_n)$ unless the skewness is zero. In fact comparison of $H_n^*(y)$ and the one-term Edgeworth expansion estimator

$$H_n^{EDG}(y) = \Phi(z) + n^{-1/2}p_n^{(1)}(z \mid \mathbf{y}_n)\phi(z)$$

where $z = \hat{\sigma}_n^{-1} y$, and $p_n^{(1)}(z \mid \mathbf{y}_n)$ is the polynomial $p^{(1)}(z)$ with the moments of $F(.)$ replaced by the sample moments calculated from $\mathbf{y}_n$, shows that both have the same asymptotic mean square error. If we considered studentized versions the bootstrap does even better. Under appropriate moment conditions both $\check{H}_n(y)$ and $\tilde{H}_n^*(y)$ have two-term Edgeworth expansions and a comparison of these shows that $\tilde{H}_n^*(y)$ has a smaller asymptotic mean square error than the corresponding one-term Edgeworth estimator, though they have the same convergence rate. The normal, bootstrap and one-term Edgeworth expansion estimators of the standardized distribution have been compared by Hall (1988) using the criterion of asymptotic relative error. When $E(|Y|^3) < \infty$ then the bootstrap does better than the normal approximation, however it may or may not do better than the one term Edgeworth expansion (see Shao and Tu, 1995). When $E(|Y|^3) = \infty$, the situation is more complicated and depends on the tail behaviour of $F(.)$. When the tail is thin the bootstrap can be worse than the normal approximation. In estimating tail behaviour the bootstrap is comparable to the one term Edgeworth expansion except in the extreme of the tail.

3.3 Asymptotic Accuracy of Confidence Intervals

The above analysis focuses on distribution functions, and does not give the whole picture. It is helpful to consider also the coverage accuracy of confidence intervals. We shall write the basic confidence limit that we seek as $\hat{\eta}_\alpha$ defined by

$$\Pr(\eta \leq \hat{\eta}_\alpha) = \alpha$$

and the normal, basic, percentile, studentized bootstrap, BC and BCa approximations as $\hat{\eta}_\alpha^{Norm}$, $\hat{\eta}_\alpha^{Boot}$, $\hat{\eta}_{\alpha,}^{Per}$ $\hat{\eta}_\alpha^{Stu}$, $\hat{\eta}_\alpha^{BC}$ and $\hat{\eta}_\alpha^{BCa}$ respectively. We summarise the results given in Shao and Tu (1995). These apply when the parameter of interest is a smooth function of the mean, $\eta = \eta(\bar{y}_n)$. Then an analysis analogous to that used for EDF's can be carried out, but now relying on an expansion of the quantile, that is the inverse of Edgeworth series, called the *Cornish-Fisher expansion*. Let $\Phi(z_\alpha) = \alpha$. Then, under appropriate moment conditions $\hat{\eta}_\alpha$ has an expansion of the form

$$q(\alpha) = z_\alpha + n^{-1/2} q^{(1)}(z_\alpha) + n^{-1} q^{(2)}(z_\alpha) + o(n^{-1})$$

where comparison with the Edgeworth exapnasion shows that

$$q^{(1)}(y) = -p^{(1)}(y)$$

and

$$q^{(2)}(y) = p^{(1)}(y)p^{(1)\prime}(y) - \frac{1}{2}p^{(1)}(y)^2 - p^{(2)}(y).$$

Under appropriate conditions we find that analogous expansions exist for the quantile approximations listed above. We find in particular that

$$\hat{\eta}_\alpha^{Boot} - \hat{\eta}_\alpha = O_p(n^{-1})$$

and in general

$$\Pr(\hat{\eta}_\alpha^{Boot} \leq \eta) = 1 - \alpha + O(n^{-1/2}).$$

However the two tailed version is second-order accurate:

$$\Pr(\hat{\eta}_\alpha^{Boot} \leq \eta \leq \hat{\eta}_{1-\alpha}^{Boot}) = 1 - 2\alpha + O(n^{-1}).$$

For symmetric distributions, these results apply to the percentile approximation as well, for example:

$$\hat{\eta}_\alpha^{Per} - \hat{\eta}_\alpha = O_p(n^{-1}).$$

The bootstrap BC method turns out to perform no better than the percentile limit, in terms of convergence rate, but the constant factor is smaller so it is marginally to be preferred. Studentization is definitely better with

$\hat{\eta}_\alpha^{Stu} - \hat{\eta}_\alpha = O_p(n^{-3/2})$ and $\hat{\eta}_\alpha^{BCa} - \hat{\eta}_\alpha = O_p(n^{-3/2})$

and

$$\begin{aligned} \Pr(\hat{\eta}_\alpha^{Stu} &\leq \eta) = 1 - \alpha + O(n^{-1}) \\ &\text{and} \\ \Pr(\hat{\eta}_\alpha^{BCa} &\leq \eta) = 1 - \alpha + O(n^{-1}). \end{aligned}$$

It follows that the two-sided intervals for both limits are also both second-order accurate. Finally we note that

$$\hat{\eta}_\alpha^{Boot} - \hat{\eta}_\alpha^{Norm} = O_p(n^{-1})$$

so that the usual normal approximation and the basic bootstrap behave similarly.

3.4 Failure of Bootstrapping

It should be clear from the previous three subsections that, for bootstrapping to work well, regularity conditions are required on both the distribution F and also on the statistic of interest. More explicitly, bootstrapping is sensitive to the tail behaviour of F; convergence of H_n^* usually requires moment conditions on F that are more stringent than those needed for convergence of H_n. Also the statistic $s(\mathbf{y})$ has to be suitably smooth in an appropriate sense. Finally it is possible for convergence of the bootstrap to be sensitive to the method used in carrying out the bootstrapping. An example of the first situation is inconsistency of $s(\mathbf{y})$, when it is simply the variance, even when the asymptotic variance is finite (see Ghosh *et al.*, 1984). This can occur if $F(.)$ is fat-tailed and does not have appropriate moments. The problem then arises because the bootstrap $s(\mathbf{y}^*)$ can take exceptionally large values. An example of the second situation is where $\mathbf{y}$ is a random sample, from $U(0, b)$ say, and we wish to consider $y_{(n)}$, the largest order statistic. Then a

natural statistic to consider is $s(\mathbf{y}) = n(b - y_{(n)})/b$, which has a limiting standard exponential distribution as $n \to \infty$. The bootstrap version is then $s(\mathbf{y}^* \mid \mathbf{y}) = n(y_{(n)} - y^*_{(n)})/y_{(n)}$. But

$$\begin{aligned} P^*(s(\mathbf{y}^* \mid \mathbf{y}) &= 0) = P^*(y^*_{(n)} = y_{(n)}) \\ &= 1 - (1 - n^{-1})^n \to 1 - e^{-1}. \end{aligned}$$

Thus H^*_n does not tend to H_n as $n \to \infty$. This result applies to any given order statistic $y_{(n-k)}$ where k is fixed. However the problem does not arise for a given quantile $y_{(pn)}$ where p is fixed with $0 < p < 1$. We shall see an example of the last situation, where convergence depends not on the problem but on the bootstraaping method employed, when we consider model selection in Section ().

4 Monte-Carlo/Simulation Models

4.1 Direct Models

Monte-Carlo or simulation models are *direct models* in the sense that they attempt to mimic the behaviour of an actual system by simulating the different objects of a system and the (typically) dynamic relationships between them. (The two names are interchangeable, but Monte-Carlo is the name more favoured by statisticians, whereas simulation is more used by the operational researchers.) A simple example is a single server queue we try to capture the actual arrival patterns of customers and how they are then processed by the server. The quantities that we try to analyse are summary characteristics such as average queue length and waiting times. We can therefore think of a set of n simulation runs as yielding observations

$$y_j = y(\mathbf{u}_j, \mathbf{v}_j, \hat{\boldsymbol{\theta}}(\mathbf{w}), \mathbf{x}_j) \quad j = 1, 2, ..., n. \qquad (19)$$

where the y_j depend on a number of quantities that we now explain. Firstly $\mathbf{u}_j$ denotes the stream of

uniformly distributed $U(0,1)$ random numbers used in the jth run. Typically the uniforms are not used directly, but are transformed into random variables drawn from distributions other than the uniform. Next comes $\mathbf{v}_j$. This represents a sequence of inputs that are random, but that has been generated independently of the simulation model. A typical instance is where $\mathbf{v}_j$ comprises sampled real observations taken from some real system, separate from the system being modelled, but on which y_j depends. Such a sampled real process is sometimes called a *trace*. We shall simply think of $\mathbf{v}_j$ as being just a sample of observations. In addition there may be further quantities which may affect y. These are denoted by $\mathbf{x}$ and $\boldsymbol{\theta}$. There is no essential difference between $\mathbf{x}$ and $\boldsymbol{\theta}$ in the way that they influence y. They are simply variables on which y depends. However we make a distinction in supposing that $\mathbf{x}$ are *decision variables*, i.e. quantities that can be selected by the simulator. However we also include in $\mathbf{x}$ parameters whose values are known, and that might affect y, but which are not selectable

by the simulator. The reason for adopting this convention is that it then allows us to assume $\boldsymbol{\theta}$ to be those parameter values, whose values are not known, and so have to be estimated. We will therefore assume that in addition to $\mathbf{v}$, there exists input data $\mathbf{w}$ that is used exclusively to estimate $\boldsymbol{\theta}$ and it is the estimated values, $\hat{\boldsymbol{\theta}}(\mathbf{w})$, that are used in the simulation runs. A simple example is a simulation model of a multiserver queue, where y is the observed average queue length over a given period, $\mathbf{v}$ might be a previously observed set of interarrival times, $\mathbf{x}$ might be the (scalar) number of servers and $\boldsymbol{\theta}$ the service rates of servers. Here we treat $\mathbf{x}$ as being selectable by the simulator so that it is a design variable, $\boldsymbol{\theta}$ may not be known and have to be estimated from available sampled real service times, $\mathbf{w}$. Resampling might already feature in carrying out the runs yielding the y_j of (19). For example variability in the $\mathbf{v}$ will contribute to the variability of y. We can therefore automatically build this effect into the simulation runs simply by using resampling from an initial given sequence $\mathbf{v}$ to construct resampled sequences $\mathbf{v}_j^*$ for use in the actual

runs. Similarly in the case (19), because $\hat{\boldsymbol{\theta}}(\mathbf{w})$ is estimated, we can allow for its randomness affecting the distribution of y by using *independent values* of $\boldsymbol{\theta}_j$ in each simulation run. When $\hat{\boldsymbol{\theta}}(\mathbf{w})$ is the ML estimate we can draw sample values of $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}_j^* \sim N(\hat{\boldsymbol{\theta}},\ I^{-1}(\hat{\boldsymbol{\theta}})) \tag{20}$$

from the asymptotic normal distribution (3) with $I(\hat{\boldsymbol{\theta}})$ calculated from (4). An alternative is to use produce bootstrap samples $\mathbf{w}_j^*$ from $\mathbf{w}$, and to use the bootstrap estimate

$$\boldsymbol{\theta}_j^* = \hat{\boldsymbol{\theta}}(\mathbf{w}_j^*) \tag{21}$$

for $\boldsymbol{\theta}$ in the jth run.

4.2 MetaModels

We now consider a very different approach where the behaviour of a system is represented by a *metamodel.* This is simply where we regard the behaviour of y as representable by a *statistical regression model*. The runs are then not represented as (19) but instead as

$$y_j = \eta(\mathbf{x}_j; \boldsymbol{\beta}) + \varepsilon_j \quad j = 1, 2, ..., n.$$

where $\eta(\mathbf{x}_j; \boldsymbol{\beta})$ is called the *regression function*. Its form is usually known, though in the problem of *model selection*, one has to choose from a number of possible competing models. The errors ε_j are assumed to be independent with $E(\varepsilon_i) = 0$. We shall consider mainly the case where all the error variances are equal $Var(\varepsilon_i) = \sigma^2$, but will consider the heteroscedastic case too. The $\boldsymbol{\beta}$ are unknown coefficients that have to be estimated. We consider linear metamodels first.

4.3 Linear Models

In the linear metamodel, observations are assumed to take the form

$$Y_j = \mathbf{x}_j^T \boldsymbol{\beta} + \varepsilon_j \quad j = 1, 2, ..., n \tag{22}$$

where the errors are assumed to be independent and identically distributed with $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$, and the $\boldsymbol{\beta}$ are unknown and have to be estimated. Let

$$\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^T, \quad \mathbf{Y} = (Y_1, ..., Y_n)^T$$

and

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

be the usual least squares estimate (equivalent to maximum likelihood when the ε_i are normally distributed). The fitted values are

$$\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y} \tag{23}$$

where

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \tag{24}$$

is the well-known known 'hat' matrix. The raw residuals are $\mathbf{r} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$. One way of resampling is to sample from these residuals as if they were the unknown errors. This gives the bootstrap sample

$$Y_j^* = \mathbf{x}_j^T \hat{\boldsymbol{\beta}} + r_j^* \quad j = 1, 2, ..., n. \tag{25}$$

We shall call this *residual sampling*. However we have that $E(\mathbf{r}) = \mathbf{0}$ and $Var(\mathbf{r}) = \sigma^2(\mathbf{I} - \mathbf{H})$. Thus a better bootstrapping method is to sample from the *adjusted residuals*

$$e_j = r_j/(1 - h_{jj})^{1/2} \tag{26}$$

where h_{jj} is the jth diagonal entry of $\mathbf{H}$ (h_{jj} commonly known as the *leverage* of the jth point). This form of model-based resampling yields bootstrap samples of form:

$$Y_j^* = \mathbf{x}_j^T \hat{\boldsymbol{\beta}} + e_j^* \quad j = 1, 2, ..., n. \tag{27a}$$

We shall call this *adjusted residual sampling*. An alternative is to use *case sampling* and to sample from the $(Y_j, \mathbf{x}_j)$ pairs directly. The main problem with this

method is that it introduces extra variation into the data. In the conventional situation $\mathbf{X}$ is regarded as fixed, but using case sampling replaces this by a bootstrap $\mathbf{X}^*$ which will be variable. This is not so much a problem when the number of parameters, p say, is small compared with n. A partial correction can often be made. For example suppose we are interested in the distribution of $\mathbf{c}^T\hat{\boldsymbol{\beta}}$ where $\mathbf{c}$ is a constant vector. Then $Var[\mathbf{c}^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] = \sigma^2\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}$. We can then study the behaviour of

$$\mathbf{c}^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/(\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}).$$

indirectly by considering the bootstrap version

$$\mathbf{c}^T(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}^*)/(\mathbf{c}^T(\mathbf{X}^{*T}\mathbf{X}^*)^{-1}\mathbf{c})$$

(where $\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}}$). A common variation is when the errors are heteroscedastic so that their variances are unequal. In this case resampling can be carried out as

$$Y_j^* = \mathbf{x}_j^T\hat{\boldsymbol{\beta}} + e_j t_j^* \quad j = 1, 2, ..., n$$

with e_i as defined in (26) and with t_j^* independent and random with $E(t_j^*) = 0$, $Var(t_j^*) = 1$. Wu (1986)

suggests sampling the t_j^* with replacement from the set

$$(r_j - \bar{r})/[n^{-1}\sum_{i=1}^{n}(r_i - \bar{r})^2]^{1/2} \quad j = 1, 2, ..., n.$$

4.4 NonLinear Models

The previous ideas apply with little change to nonlinear models. For additive errors we have

$$Y_j = \eta(\mathbf{x}_j; \boldsymbol{\beta}) + \varepsilon_j \quad j = 1, 2, ..., n. \tag{28}$$

where $\eta(\mathbf{x}_j; \boldsymbol{\beta})$ is not necessarily linear in $\boldsymbol{\beta}$. Again $\hat{\boldsymbol{\beta}}$ can be obtained by ML say and residuals formed: $r_j = Y_j - \eta(\mathbf{x}_j; \hat{\boldsymbol{\beta}}) \quad j = 1, 2, ..., n$. This time there is no obvious equivalent of standardization so the bootstrap sample is simply

$$Y_j^* = \eta(\mathbf{x}_j; \hat{\boldsymbol{\beta}}) + r_j^* \quad j = 1, 2, ..., n.$$

However the method extends easily to more complex nonlinear situations. For example it may be that we have observations $(y_j, \mathbf{x}_j)$ where

$$y_j \sim Poisson(\mu_j), \quad \mu_j = \exp(\mathbf{x}_j^T \boldsymbol{\beta}).$$

Again $\hat{\boldsymbol{\beta}}$ can be obtained by ML and a bootstrap sample can then be formed by drawing a sample as

$$y_j^* \sim Poisson(\mu_j^*), \quad \mu_j^* = \exp(\mathbf{x}_j^T \hat{\boldsymbol{\beta}}).$$

A clear introduction to resampling using *generalized linear models* is given by Hjorth (1994).

4.5 Uses of Metamodels

We end by discussing some examples of *why* we might wish to carry out bootstrap resampling of a regression metamodel. We give three examples. The first is when the regression function $\eta(\mathbf{x};\boldsymbol{\beta})$ represents a performance index and we might wish to construct a confidence interval for it at some given value of $\mathbf{x}$, $\mathbf{x}_0$ say. Suppose that we draw B sets of bootstrap observations of the form (27a)

$$\{Y_j^{(i)*} = \eta(\mathbf{x}_j;\hat{\boldsymbol{\beta}}) + e_j^{(i)*}, j = 1, 2, ..., n\},\ i = 1, 2, ..., B,$$

and from each set we calculate a bootstrap estimator $\hat{\boldsymbol{\beta}}^{(i)*}$ of $\hat{\boldsymbol{\beta}}$ using the ML method say. Then the corresponding set of bootstrap regression function values $\eta(\mathbf{x}_0, \hat{\boldsymbol{\beta}}^{(i)*})$ can be used to calculate a bootstrap confidence interval. For example we could use (15) with $\hat{\boldsymbol{\eta}}^*$ in that formula having components $\hat{\eta}_i^* = \eta(\mathbf{x}_0, \hat{\boldsymbol{\beta}}^{(i)*})$. The second example is when we wish to find an optimal setting for $\mathbf{x}$. We consider the simplest case where

$$Y = \eta(x;\boldsymbol{\beta}) + \varepsilon,$$

with x scalar and where

$$\eta(x;\boldsymbol{\beta}) = \beta_0 + \beta_1 x + \beta_2 x^2.$$

Then the optimal setting is given by $d\eta/dx = 0$, i.e. at $x_{opt} = -\beta_1/(2\beta_2)$. Using ML to estimate this yields

$$\hat{x}_{opt} = -\hat{\beta}_1/(2\hat{\beta}_2).$$

Again a simple confidence interval for the unknown x_{opt} is given by (15), only this time, with $\hat{\boldsymbol{\eta}}^*$ in that formula having components $\hat{\eta}_i^* = -\hat{\beta}_1^{(i)*}/(2\hat{\beta}_2^{(i)*})$. This is an example where bootstrapping furnishes a relatively easy answer to what in classical statistics is a problem that is not straightforward. The final example concerns the identification of important factors. Suppose we are interested in identifying those coefficients in the regression function $\eta(\mathbf{x};\boldsymbol{\beta})$ for which $|\beta_i| > \beta_{0i}$ where $\beta_{0i} > 0$ are given constants. Let $S = \{\beta_i : |\beta_i| > \beta_{0i}\}$ denote the *important set of coefficients*. The obvious estimate is to select those coefficients β_i for which $\left|\hat{\beta}_i\right| > \beta_{0i}$,i.e. $\hat{S} = \{\beta_i : \left|\hat{\beta}_i\right| > \beta_{0i}\}$. Bootstrapping is a simple way of assessing the

stability of this choice. We generate B bootstrap samples $(y_j^{(k)*}, \mathbf{x}_j^{(k)*})$ $j = 1, 2, ..., n,$ $k = 1, 2, ..., B,$ using either residual sampling or case sampling say and fit the regression model to each bootstrap sample to give bootstrap estimates $\hat{\boldsymbol{\beta}}^{(k)*},$ $k = 1, 2, ..., B.$ We then calculate $\hat{S}^{(k)*} = \{\beta_i : \left|\hat{\beta}_i^{(k)*}\right| > \beta_{0i}\},$ $k = 1, 2, ..., K.$ Now assume that B is sufficiently large so that each distinct bootstrap important set that has been obtained occurs a reasonable number of times. Then a $(1 - \alpha)$ confidence region for the unknown true important set can be constructed by selecting bootstrap important sets in decreasing order of their observed frequency of occurrence until $(1 - \alpha)100\%$ of the $\hat{S}^{(k)*}$ have been chosen. Related to identifying important factors in metamodels is the rather more difficult problem of metamodel selection. We consider this in the next section.

4.6 Metamodel Comparison and Selection-odels

Metamodel comparison and selection is a difficult subject. The main trouble is that models that we wish to compare may be of different functional complexity. The statistical properties associated with quantities derived from different models may therefore be hard to put on a common scale on which they can be compared. A reasonably satisfactory approach is based on the use of *cross-validation* technique. We suppose that the initial data has the form

$$S = \{(y_j, \mathbf{x}_j), \quad j = 1, 2, ..., n\}$$

with

$$y_j = \eta(\mathbf{x}_j, \boldsymbol{\beta}) + \varepsilon_j, \quad j = 1, 2, ..., n.$$

Suppose that our fitted regression is

$$\hat{y}(\mathbf{x}) = \eta(\mathbf{x}, \hat{\boldsymbol{\beta}}).$$

In its basic form cross-validation is used to assess how effective the fitted regression is for predicting the response at some new design point $\mathbf{x}_{new}$. Rather than

explicitly choosing such a new point, a simple idea is the *leave-one-out* method where we drop an observed point $(y_j, \mathbf{x}_j)$ from the set of all observed points, fit the metamodel to the remaining points $S_{-j} = S \setminus (y_j, \mathbf{x}_j)$ giving the fitted regression as

$$\hat{y}_{-j}(\mathbf{x}) = \eta(\mathbf{x}, \hat{\boldsymbol{\beta}}_{-j}),$$

and then look at the squared difference between the omitted y_j and the value of y at $\mathbf{x}_j$, as predicted by the fitted model; i.e.

$$(y_j - \hat{y}_{-j}(\mathbf{x}_j))^2.$$

If we do this even handedly by leaving out each observation in turn we have as an *estimate of cross-validation prediction error*:

$$\hat{L}_{CV} = n^{-1} \sum_{j=1}^{n} (y_j - \hat{y}_{-j}(\mathbf{x}_j))^2. \qquad (29)$$

It turns out that, for the linear regression model, this formula simplifies elegantly to one where we only need

comparisons with the one model fitted to all the original observations:

$$\hat{L}_{CV} = n^{-1} \sum_{j=1}^{n} \frac{(y_j - \hat{y}(\mathbf{x}_j))^2}{1 - h_{jj}}$$

where

$$\hat{y}(\mathbf{x}_j) = \mathbf{x}_j^T \hat{\boldsymbol{\beta}},$$

and h_{jj} is the jth main diagonal entry in the hat-matrix (24), as before. The distributional property of $\hat{L}_{CV}$ can be obtained in the usual way by bootstrapping to get B bootstrap samples. If we use case resampling say then

$$S^{(i)*} = \{(y_j^{(i)*}, \mathbf{x}_j^{(i)*}), \quad j = 1, 2, ..., n\}$$

where each $(y_j^{(i)*}, \mathbf{x}_j^{(i)*})$ is an observation drawn at random from the set S. We turn now to model selection. For simplicity we shall only consider the linear case where $y = \mathbf{x}^T \boldsymbol{\beta} + \varepsilon$, though with an obvious adjustment, the discussion does generalize. Let $\boldsymbol{\beta}$ be of dimension p. If we are uncertain as to which design variables are really needed in the model, then we

would have in principle up to 2^p models to choose from with $\binom{p}{q}$ (sub)models where there were precisely q covariates selected. We denote a typical submodel by M. The estimate of prediction error using (29) is

$$\hat{L}_{CV}(M) = n^{-1} \sum_{j=1}^{n} \frac{(y_j - \hat{y}_M(\mathbf{x}_j))^2}{1 - h_{Mjj}}.$$

It turns out that this measure does not work as well as it might even when n is large. (This is an example previously referred to in subsection () where inconsistency arises not in the problem itself but in the bootstrap sampling method.) A much better variant is not to leave out just one observation at a time but instead to split the observations into two: a *training* set, and an *assessment* set with respectively $n_t = n - m$ and $n_a = m$ observations in each. We shall moreover select not one but K such pairs and denote the sets as $S_{t,k}$ and $S_{a,k}$, $k = 1, 2, ..., K$. We shall write $\hat{\boldsymbol{\beta}}_{Mk}$ for the coefficients of model M fitted to the kth training set, and write $\hat{y}_{Mjk} = \mathbf{x}_{Mj}^T \hat{\boldsymbol{\beta}}_{Mk}$ for the value of

$\mathbf{x}_j^T\boldsymbol{\beta}$ predicted by this model M at $\mathbf{x}_j$. Then $\hat{L}_{CV}(M)$ becomes

$$\hat{L}_{CV}(M) = K^{-1}\sum_{k=1}^{K} m^{-1} \sum_{j\in S_{a,k}} (y_j - \hat{y}_{Mjk})^2. \quad (30)$$

We use the same set of K pairs for all models being compared. Provided $n - m \to \infty$ and $m/n \to 1$ as $n \to \infty$ then selecting M to minimize (30) will yield a consistent estimator for the correct model. When n is small it may not be possible to select m large enough in which case Davison and Hinkley (1997) suggest taking $m/n \simeq 2/3$. The variability of $\hat{L}_{CV}(M)$ can be estimated by bootstrapping in the usual way.

5 Bootstrap Comparisons

In this Section we consider problems where there are a number of different samples to compare. We suppose that we have m data sets

$$\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{i,n_i}), \quad i = 1, 2, ..., m \tag{31}$$

the ith being of size n_i runs. This situation covers two different scenarios in particular. The first is where we are considering the question: Does the output of the simulation model accurately represent the output of the system being modelled? Here we have just two data sets, so that $m = 2$, and one data set comprises observations of a real system whilst the other comprises the observed output from a simulation model. Thus this is a question of validation. The second is when we have a number of different simulation models to assess. Here validation against real data is not the issue. The different models simply represent different systems, and their validity is not in doubt. Instead we are just interested in assessing how differently the systems behave, as represented by the output of their

simulation models. Typically we might be interested in which system produces the greatest average output, or least average delay. Alternatively we might wish to compare output variability. A similar methodology can be applied to either problem. We discuss this in the remainder of this section.

5.1 Goodness-of-Fit and Validation

A discussion of validation involving trace-driven simulation is described in Kleijnen *et al.* (2000, 2001). The basic idea used there can be generalized. We treat the problem as essentially one of goodness-of-fit. A simple starting point is where we have n simulation outputs y_j, $j = 1, 2, ..., n$, to compare with m observations y_j^{Real}, $j = 1, 2, ..., m$ of a real system and we wish to know if the two samples are identically distributed. A good goodness-of-fit statistic is the two sample Cramer - von Mises statistic (see Anderson, 1962):

$$W^2 = \int [F_n(y) - F_m^{Real}(y)]^2 dH_{n+m}(y) \qquad (32)$$

where F_n, F_m^{Real} and H_{n+m} are respectively the EDF's of the simulated, real and combined samples. The asymptotic distribution of W^2 is known when the samples are drawn from continuous distributions, but it is easy to set up a bootstrap version of the test that can handle discrete distributions just as well. We can use a

bootstrapping method as follows. We obtain B pairs of bootstrap samples $(\mathbf{y}^{(i)*}, \mathbf{y}^{Real(i)*})$, $i = 1, 2, ..., B$, each sample in the pair being obtained by resampling with replacement from $\mathbf{y}$ and $\mathbf{y}^{Real}$ combined but with $\mathbf{y}^{(i)*}$ the same sample size as $\mathbf{y}$, and $\mathbf{y}^{Real(i)*}$ the same sample size as $\mathbf{y}^{Real}$. Denote the EDFs of the samples of each pair by $F_n^{(i)*}$ and $F_n^{Real(i)*}$, $i = 1, 2, ..., B$. We can now calculate B bootstrap two sample Cramer - von Mises statistics from each pair:

$$W^{(i)*2} = \int [F_n^{(i)*}(y) - F_m^{Real(i)*}(y)]^2 dH_{n+m}^{(i)*}(y). \tag{33}$$

Under the null assumption that $\mathbf{y}$ and $\mathbf{y}^{Real}$ are drawn from the same distribution it follows that each $W^{(i)*2}$ is just a bootstrap version of W^2. We can thus calculate a critical p-value from the EDF of the $W^{(i)*2}$, $i = 1, 2, ..., B$. The null hypothesis that $\mathbf{y}$ and $\mathbf{y}^{Real}$ are drawn from the same distribution can be tested simply by checking if the original W^2 exceeds this p-value or not. The validation method just described

is easily extended to allow the representational accuracy of a number of different competing simulation models to be compared against (the same) real data. We order the Cramer von-Mises goodness of fit statistics (32) obtained from comparing the real data with the simulated output of each of the models. We can assess the reliability of the comparison by bootstrapping as in (33) to obtain bootstrap distributions of each such Cramer von-Mises statistic and looking at the degree of overlap of amongst these distributions. When the W^2 statistic shows the data samples $\mathbf{y}_j$ to be significantly different, there is an interesting decomposition of W^2 that allows the nature of the differences to be more clearly identified . Durbin and Knott (1972) show that in the one sample case where $F_n(.)$ is being compared with a uniform null distribution, so that $W^2 = n \int [F_n(y) - y]^2 dy$, then W^2 has the orthogonal representation

$$W^2 = \sum_{j=1}^{\infty} (j\pi)^{-2} z_{nj}^2$$

where the z_{nj} are scaled versions of the coefficients in the Fourier decomposition of the process$\sqrt{n}\,(F_n(y)-y)$. The z_{nj} are stochastic of course, depending on $F_n(.)$, but they are independent and, under the null, have identical distributions. It can be shown that z_{n1}, z_{n2} and z_{n3} are dependent essentially exclusively on deviations respectively of the mean, variance and skewness of $F()$ from that of the null. In other words these components z_{n1}, z_{n2} and z_{n3} can be used to provide convenient statistical tests for these differences. Cheng and Jones (2000, 2004) describe a generalisation of the results of Durbin and Knott (1972), which they term *EDFIT statistics,* suitable for comparison of data of the form (31) arising in the simulation context. Their formulation uses ranks rather than the original. Though this leads to some loss of power the method does then have the advantage of enabling critical values of tests to be easily obtained by bootstrapping, and moreover in an exact way.

5.2 Comparison of Different Systems

We now consider the situation of (31) where the data sets are outputs from different simulation models only, and no comparison is made with real data. Here we are interested simply in making comparisons between different models. The discussion of the previous subsection goes through with no essential change. The only difference is that all samples have the same status; there is no sample being singled out as special by being drawn from a real system. The EDFIT approach can therefore be used directly, with bootstrapping providing critical null values. A more direct approach, not using goodness of fit ideas, is possible. We focus on examining differences between the means, $\bar{y}_i$, $i = 1, 2, .., m$, of the m samples of (31). Comparison of other sample statistics be handled in the same way. If we suppose that largest is best we can rank the means as $\bar{y}_{(1)} > \bar{y}_{(2)} > ... > \bar{y}_{(m)}$. The question then is how stable is this order? Bootstrapping provides an easy answer. We generate B bootstrap sets

of observations

$$\begin{aligned} \mathbf{y}_i^{(k)*} &= (y_{i1}^{(k)*}, y_{i2}^{(k)*}, ..., y_{i,n_i}^{(k)*}), \\ i &= 1, 2, ..., m, \quad k = 1, 2, ..., B \end{aligned} \tag{34}$$

where each sample $\mathbf{y}_i^{(k)*},\ k = 1, 2, ..., B$, is obtained by sampling with replacement from $\mathbf{y}_i$. We then order the means in each bootstrapped set: $\bar{y}_{(1)}^{(k)*} > \bar{y}_{(2)}^{(k)*} > ... > \bar{y}_{(m)}^{(k)*},\ k = 1, 2, ..., B$. The frequency count of how many times the mean of a given sample comes out on top is a measure its relative merit. A more comprehensive picture is provided by setting up a two-way table showing the number of times $\bar{y}_i^{(k)*} > \bar{y}_j^{(k)*}$ out of the B bootstrapped sets, for each possible pair $1 \leq i,\, j \leq m$. The parametric form of the bootstrapping procedure is equally easy to implement, though it is not clear there is much advantage to be had in doing so. Suppose for example that the ith sample $\mathbf{y}_i$ is drawn from the distribution $F^{(i)}(., \boldsymbol{\theta}_i)$. Usually the $F^{(i)}(., \boldsymbol{\theta}_i)$ will have the same functional form, for example all normal distributions, but the procedure works equally well when the $F^{(i)}$ are functionally

different. Let the mean of the $F^{(i)}$ distribution be $\mu^{(i)}(\boldsymbol{\theta}_i)$. Then we can estimate, by ML say, the parameters of each distribution from their corresponding sample. Let these estimates be $\hat{\boldsymbol{\theta}}_i$. We can then carry out parametric bootstrapping by forming (34) but this time with each sample $\mathbf{y}_i^{(k)*}$, $k = 1, 2, ..., B$, obtained by sampling from the fitted distribution $F^{(i)}(., \hat{\boldsymbol{\theta}}_i)$. The analysis then proceeds as before.

6 Bayesian Models

Much of the current popularity of Bayesian methods comes about because increased computing power makes possible Bayesian analysis by numerical procedures, most notably Markov Chain Monte Carlo (MCMC). See Gilks et al. (1996).This allows numerical sampling to be carried out in much the same way as is done in bootstrap and other resampling procedures. We indicate the commonality by considering just one or two situations where Bayesian methods and resampling overlap. We again focus on a random sample $\mathbf{w} = (w_1, w_2, ..., w_n)$ drawn from a distribution $F(w, \boldsymbol{\theta})$. As in the situation where we considered ML estimation, the form of F is known but it depends on a vector $\boldsymbol{\theta}$ of parameters. In the ML estimation situation $\boldsymbol{\theta}_0$,the true value of $\boldsymbol{\theta}$, was assumed unknown. In the Bayesian case a prior distribution is assumed known. We consider just the continuous case and denote the pdf of the prior by $\pi(\boldsymbol{\theta})$. The main step in Bayesian analysis is to construct the posterior distribution $\pi(\boldsymbol{\theta}$

$\mid \mathbf{w})$ which shows how the sample $\mathbf{w}$, which depends $\boldsymbol{\theta}$, has modified the prior. The Bayesian formula is

$$\pi(\boldsymbol{\theta} \mid \mathbf{w}) = \frac{p(\mathbf{w} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int p(\mathbf{w} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}. \qquad (35a)$$

The difficult part is the evaluation of the normalizing integral $\int p(\mathbf{w} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$. MCMC has proved remarkably successful in providing a powerful numerical means of doing this. In the context of simulation a Bayesian approach is useful in assessing uncertainty concerning input parameter values used in a simulation model. We can adopt the viewpoint given in (19) and regard the runs as taking the form:

$$y_j = y(\mathbf{u}_j, \mathbf{v}_j, \boldsymbol{\theta}, \mathbf{x}_j) \quad j = 1, 2, ..., n. \qquad (36)$$

only now $\boldsymbol{\theta}$ is assumed to have a Bayesian prior distribution $\pi(\boldsymbol{\theta})$. Note that in this situation we do *not* have the equivalent of data $\mathbf{w}$ from which to calculate a posterior distribution for $\boldsymbol{\theta}$. What we can do is to treat the prior $\pi(\boldsymbol{\theta})$ as *inducing a prior distribution* on $y = y(\mathbf{u}, \mathbf{v}, \boldsymbol{\theta}, \mathbf{x})$. In this sense we can think of y itself as having a prior which can then be estimated

by sampling $\boldsymbol{\theta}$ from its prior $\pi(\boldsymbol{\theta})$, yielding values $\boldsymbol{\theta}_j$ $j = 1, 2, ..., n$, and then running the model to produce a set of observations

$$y_j = y(\mathbf{u}_j, \mathbf{v}_j, \boldsymbol{\theta}_j, \mathbf{x}_j) \quad j = 1, 2, ..., n. \tag{37}$$

The EDF of these y_j then estimates this prior distribution of y. This process clearly is the analogue to the classical situation where $\boldsymbol{\theta}$ is estimated from data $\mathbf{w}$, as is supposed in (19), and we then allowed for this variability in the y_j by replacing $\hat{\boldsymbol{\theta}}(\mathbf{w})$ by $\boldsymbol{\theta}_j^*$ calculated from (20) or (21). More interestingly we can take this a step further if there are real observations available of the y process itself. We do not attempt a fully general formulation but give an example to indicate the possibilities. Suppose that

$$y_j^{Real} = y^{Real}(\mathbf{x}_j), \qquad j = 1, 2, ..., n$$

comprise n observations on the real system being modelled by the simulation. The precise relationship between the simulation output $y = y(\mathbf{u}, \mathbf{v}, \boldsymbol{\theta}, \mathbf{x})$ and

$y^{Real}(\mathbf{x})$ is not known. However we might assume that

$$y(\mathbf{u}, \mathbf{v}, \boldsymbol{\theta}, \mathbf{x}) \sim N(y^{Real}(\mathbf{x}),\ \sigma^2)$$

where σ^2 is an additional parameter that will also be treated as Bayesian in the sense of having a prior. We can express our great prior uncertainty about σ by assuming a *reference prior distribution*, $\rho(\sigma)$, for it. Then, by (35a), the posterior distribution of $(\boldsymbol{\theta}, \sigma)$ is proportional to

$$\sigma^{-n}\pi(\boldsymbol{\theta})\rho(\sigma)\exp\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}[y(\mathbf{u}, \mathbf{v}, \boldsymbol{\theta}, \mathbf{x}_j) - y^{Real}(\mathbf{x}_j)]^2\}.$$

The posterior distribution can thus be obtained by MCMC methods for example. An interesting application of this problem occurs in epidemiological modelling. Suppose that y is a measure of the progress of an epidemic that is dependent on factors $\boldsymbol{\theta}$ for which there are previous measurements or for which there is expert information. A Bayesian approach is then a natural way of incorporating this prior information.

We need also to have a good epidemic model for producing a simulated y. Thus reverse use of simulation as indicated above has allowed this prior information to be updated.

7 Time Series Output

Bootstrapping of time series is a well studied problem. In simulation the most likely use of such procedures is to generate correlated input for a model. As usual the parametric form is relatively easy to explain and implement and we discuss this first.

7.1 Residual Sampling

We consider the case where the time-series is an autoregressive model. Here the residual sampling method used to construct a bootstrap metamodel applies with little change. Suppose we have the autorgressive model

$$Y_t = \sum_{j=1}^{p} a_j Y_{t-j} + \varepsilon_t$$

where the ε_t are independently and identically distributed, commonly called the *innovations.* Suppose that $y_1, y_2, ..., y_n$ are a series drawn from this model. Then we can estimate the a_j by least squares say,

yielding the estimates $\hat{a}_j$, $j = 1, 2, ..., p$, and form the residuals

$$r_t = y_t - \sum_{j=1}^{p} \hat{a}_j y_{t-j}, \quad t = p+1, p+2, ..., n.$$

We can then form the bootstrap time-series as

$$y_t^* = \sum_{j=1}^{p} \hat{a}_j y_{t-j}^* + r_t^*, \quad t = 1, 2, ..., n$$

by sampling the r_t^* from the EDF of the residuals $\{r_t\}$. We need $y_1^*, y_2^*, ..., y_p^*$ to initiate the process, but if we assume that the observed series is stationary, it is probably easiest to simply initiate the series with some arbitrary starting values, possibly the original $y_1, y_2, ..., y_p$, then run the bootstrapping sufficiently long until initial effect are negligible and collect the actual y_t^* from that point on. Freedman (1984) gives conditions for the asymptotic validity of this procedure, Basawa *et al.* (1989) extending these results to the case of nonlinear time-series.

7.2 Block Sampling

For time-series the analogue of case sampling is *block sampling.* We cannot sample individual observations y_t because this loses the correlation between observations. If the series is long enough then we can take $n = bl$ and think of the series as comprising b blocks each of length l. We write the ith block as $\mathbf{y}_i = (y_{l(i-1)+1}, y_{l(i-1)+2}, ..., y_{(li)})$ $i = 1, 2, ..., b$. Bootstrapping is done by sampling blocks with replacement from this set of b blocks, retaining the order of the observations in each block when writing down the individual observations of the bootstrapped series. A balance needs to be struck between having block lengths long enough to retain the correlation properties between neighboring observations, and having enough blocks to measure the inherent variation of the series. A typical compromise is to use say $b = l = n^{1/2}$ so that both quantities tend to infinity as $n \to \infty$. A major weakness of block sampling is the loss of correlation incurred by the random sampling of blocks. This loss of correlation is called

whitening. It is especially serious when the statistic of interest involves correlations of high lag. The crude block sampling just described may be quite ineffective if the size of blocks is not large enough, because calculation involves quantities which straddle blocks and which are then not sufficiently correlated because of the whitening. There are many variants of block sampling aimed at reducing the effect of whitening in specific situations. A good example is estimation of the lag m covariance

$$c_m = \frac{1}{n-m} \sum_{t=1}^{n-m} (y_t - \bar{y})(y_{t+m} - \bar{y}).$$

Here we can define a two-dimensional process

$$\mathbf{z}_t = \begin{pmatrix} z_{1t} \\ z_{2t} \end{pmatrix} = \begin{pmatrix} y_t \\ y_{t+m} \end{pmatrix}, \quad t = 1, 2, .., n-m$$

with $\bar{z}_i = (n-m)^{-1} \sum_{t=1}^{n-m} z_{it}$ and think of c_1 as a statistic of this process

$$c_1 = \frac{1}{n-1} \sum_{t=1}^{n-1} (z_{1t} - \bar{z}_1)(z_{2t} - \bar{z}_2).$$

We can then obtain bootstrap $\mathbf{z}_t^*$ by sampling with replacement from the set $\{\mathbf{z}_t\}$.The bootstrap lag m covariance then clearly substantially retains the covariance of the original series as we have, in effect, bootstrap sampled the terms $(y_t - \bar{y})(y_{t+m} - \bar{y})$ appearing in the formula giving c_m. Generalizations of this technique are known as *block of blocks sampling*.

7.3 Spectral resampling

Residual and block sampling are time domain techniques. An alternative approach is to sample in the frequency domain. A big advantage is that spectral increments are uncorrelated, and for Gaussian processes this strengthens to independent increments. Suppose we have $n = 2m + 1$ observations

$$y_t,\ t = -m, -m+1, ..., m-1, m$$

for which there is a continuous spectral density $S(\omega)$ and define the frequencies $\omega_k = 2\pi k/n$, $-m \leq k \leq m$. Then the observations have the spectral representation

$$y_t = \sum_{k=-m}^{m} a_k e^{i\omega_k t}$$

where

$$a_k = n^{-1} \sum_{t=-m}^{m} y_t e^{-i\omega_k t}.$$

In this section $i = \sqrt{-1}$. For a Gaussian process the real and purely imaginary components of the a_k are

independent, and the a_k are asymptotically so. Norgaard (1992) gives two possible ways of obtaining bootstrap samples of the a_k. The simplest version is to draw a_k^* at random from one of the twelve elements

$$(\pm a_{k-1},\ \pm a_k,\ \pm a_{k+1},\ \pm ia_{k-1},\ \pm ia_k,\ \pm ia_{k+1}) \tag{38}$$

when $0 < k < n$ and to draw a_n^* at random from one of the twelve elements

$$(\pm a_{n-1},\ \pm a_n,\ \pm a_n,\ \pm ia_{n-1},\ \pm ia_n,\ \pm ia_n). \tag{39}$$

(note the repetition of some elements in this latter case). In both cases we set

$$a_{-k}^* = \overline{a_k^*}$$

i.e. a_{-k}^* is the complex conjugate of a_k^*. The value of a_0^* needs to be real. The simplest case is when $E(Y_t) = 0$ when we can select a_0^* at random as one of the four elements $(\pm a_0,\ \pm|a_1|)$. A variant which seems to be significantly superior is called the *extended circle* (EC) method by Norgaard (1992). Here we select $|a|_k^*$ as a random element of the set

$(|a_{k-1}|,\ |a_k|,\ |a_{k+1}|)$ and then give this a random spin in both the real and imaginary directions:

$$a_k^* = |a|_k^*(\cos\phi_{1k} + i\sin\phi_{2k}),\ k = 1, 2, ..., n$$

where ϕ_{1k} and ϕ_{2k} are independent $U(-\pi, \pi)$ angles. We have $a^*_{-k} = \overline{a_k^*}$ as before.

8 Final Comment

There are many topics not covered in this appendix that could well have been included. We have for instance, not touched on the closely related technique of jackknifing. Other topics we have made only cursory mention, such as cross validation and when bootstrapping fails.

References

[1] Anderson, T. W. 1962. On the distribution of the two-sample Cramer-von Mises criterion. *Ann. of Math. Statist.,* **33,** 1148-1159.

[2] Bickel, P. J. and D. A. Freedman. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.*, **9**, 1196-1217.

[3] Cheng, R. C. H. and O. D. Jones. 2000. Analysis of simulation factorial experiments by EDF resample statistics. *Proceedings of the 2000 Winter Simulation Conference*, eds J. A. Joines, R. Barton, P. Fishwick, K. Kang. Piscataway, New Jersey: IEEE, 697-703.

[4] Cheng, R. C. H. and Jones, O.D. 2004. Analysis of distributions in factorial experiments. *Statistica Sinica.* To appear.

[5] Cheng, R. C. H. 1995. Bootstrap methods in computer simulation experiments. In Proceedings of 1995 Winter Simulation Conference, eds W. R. Lilegdon, D. Goldsman, C. Alexopoulos and K. Kang. Piscataway, New Jersey: IEEE, 171-177.

[6] Chernick, M. R. 1999. *Bootstrap Methods.* New York: John Wiley.

[7] Cox, D. R. and D. V. Hinkley. 1974. *Theoretical Statistics.* London: Chapman & Hall.

[8] Davison, A. C., D. V. Hinkley and E. Schechtman. 1986. Efficient bootstrap simulations. *Biometrika*, **73**, 555-566.

[9] Davison, A. C. and D. V. Hinkley. 1997. *Bootstrap Methods and Their Application.* Cambridge, England: Cambridge University Press.

[10] Durbin, J. and M. Knott. 1972. Components of Cramer-von Mises statistics I. *J. Royal Statist. Soc.,* **34,** 290-307.

[11] Efron, B. 1979. Bootstrap methods: another look at the jackknife. *Ann. Statist.*, **7**, 1-26.

[12] Efron, B. 1987. Better bootstrap confidence intervals. *J. Amer. Statist Assoc.* **82**, 171-185.

[13] Efron, B. and R. J. Tibshirani. 1993. *An Introduction to the Bootstrap.* New York and London: Chapman and Hall.

[14] Gilks, W. R., S. Richardson and D. J. Spiegelhalter. 1996. *Markov Chain Monte Carlo in Practice.* London: Chapman & Hall.

[15] Hall, P. 1988. Rate of convergence in bootstrap approximations. *Ann. Prob.*, **16**, 1165-1685.

[16] Hall, P.and Pittelkow, Y.E. 1990, "Simultaneous bootstrap confidence bands in regression." J. Statist. Comput. Simul. 37, 99-113.

[17] Hall, P. 1992. *The Bootstrap and Edgeworth Expansion.* New York: Springer-Verlag.

[18] Hjorth, J. S. U. 1994. *Computer Intensive Statistical Methods.* London: Chapman & Hall.

[19] Johns, M. V. Jr. 1988. Importance sampling for bootstrap confidence itervals. *J. Amer. Statist. Assoc.*, **83**, 709-714.

[20] Kleijnen, J.P.C., R. C. H. Cheng and B. Bettonvil. 2000. Validation of trace-driven simulation models: more on bootstrap tests. In *Proceedings of the 2000 Winter Simulation Conference*, eds J. A. Joines, R. R. Barton, K. Kang and P. A. Fishwick, IEEE Piscataway, New Jersey, 882- 892

[21] Kleijnen, J. P. C., R. C. H. Cheng and B. Bettonvil. 2001 Validation of Trace Driven Simulation Models: Bootstrap Tests, *Management Science*, **47**, 1533-1538.

[22] Norgaard, A. 1992. Resampling stochastic processes using a bootstrap approach. In Bootstrapping and Related Topics. *Lecture Notes in Economics and Mathematical Systems*, Springer, **376**, 181-185.

[23] Serfling, R. J. 1980. *Approximation Theorems of Mathematical Statistics.* New York: Wiley.

[24] Shao, J. and D. Tu. 1995. *The Jackknife and the Bootstrap.* New York: Springer.

[25] Singh, K. 1981. On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist.*, **9**, 1187-1195.

[26] van Beeck, P. 1972. An application of Fourier methods to the problem of sharpening the Berry-Esséen inequality. *Z. Wahrscheinlichkeitstheorie und. Verw. Gebiete*, **15**, 279-290.

[27] von Mises, R. 1947. On the asymptotic distribution of differentiable statistical functions. *Ann. Math. Statist.*, **18**, 309-348.

[28] Wilks, S. S. 1962. *Mathematical Statistics.* New York: Wiley.

[29] Wu, C. F. J. 1986. Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.*, **14**, 1261-1295.