Bootstrap Diagnostics

A. J. Canty[1], A. C. Davison[2], D. V. Hinkley[3], and V. Ventura[4]

1: *Department of Mathematics and Statistics, Concordia University, Montréal, Quebec, Canada H3G 1M8* `canty@discrete.concordia.ca`

2: *Department of Mathematics, Swiss Federal Institute of Technology, 1015 Lausanne, Switzerland* `Anthony.Davison@epfl.ch`

3: *Department of Statistics and Applied Probability, University of California, Santa Barbara, CA 93106–3110, USA* `hinkley@pstat.ucsb.edu`

4: *Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, USA* `vventura@stat.cmu.edu`

September 19, 2000

Summary

Bootstrap diagnostics are used to assess reliability of bootstrap calculations, and may suggest useful modified calculations when these are possible. Concern focuses on susceptibility to peculiarities in data, incorrectness of resampling model, incorrect use of resampling simulation output, and inherent inaccuracy of the bootstrap approach. The last involves issues such as inconsistency of bootstrap method, order of correctness of a consistent bootstrap method, and approximate pivotality. In this paper we attempt to address these problems, to provide workable diagnostic methods, and where necessary to provide fast and simple ways to effect the necessary computations.

Some key words: Blocking; Bootstrap recycling; Breakdown; Count data; Extreme values; Importance resampling; Inconsistency; Jackknife-after-bootstrap; Linear approximation; Long-tailed distributions; Nested bootstrap; Outlier; Pareto distribution; Periodogram; Pivot; Resampling model; Robust estimate; Shrinkage estimator; Smoothers; Stein estimator; Stratification; Superefficiency; Tiling; Time series; Variance function; Wavelet regression.

1 Introduction

Over the past decade bootstrap methods have become increasingly widely used in statistical applications. The siren-call of solutions apparently tailor-made for a wide range of problems, together with the vastly increased power of desktop computers, has seduced many researchers into using resampling methods to estimate significance probabilities, set confidence intervals, and so forth. But as with any statistical procedure, it is important to be clear about the assumptions that underlie the validity and accuracy of a bootstrap calculation. The purpose of this paper is to discuss methods for checking some of these assumptions, and ways of using the methods to modify bootstrap calculations if possible.

One simple way to describe the bootstrap methods in general is to say that they are applications of the *bootstrap substitution principle*, which says that if we want to know the probability distribution of the quantity $U \equiv u(Y, F)$ when $Y = (Y_1, \ldots, Y_n)$ is randomly sampled from F, then we replace F in the probability calculation by a fitted model $\hat{F}$, i.e. we make the approximation

$$\Pr\{u(Y,F) \leq u \mid F\} \approx \Pr\{u(Y^*, \hat{F}) \leq u \mid \hat{F}\}. \tag{1}$$

Here the superscript * is used to distinguish random variables (and related quantities) sampled from a probability model that has been fitted to the data. In the simplest applications $u(Y, F) = T - \theta$, where T is an estimator of the population parameter $\theta = t(F)$. More complicated situations will involve transformations of T or studentized forms of T. Usually $T = t(\tilde{F})$ where $\tilde{F}$ denotes the empirical distribution function (EDF) of the data values. Corresponding approximations are made for characteristics of T such as bias and variance.

The fitted model $\hat{F}$ in (1) may correspond either to a parametric estimate $F_{\hat{\psi}}$, where the full parameter vector ψ includes the parameter of interest θ, or to the EDF $\tilde{F}$, which assigns equal probabilities to each case in this simple setting. In more complicated settings, such as regression or time series, $\hat{F}$ may correspond to a semiparametric model, for example a parametric regression fit with additive errors whose distribution is estimated by the EDF of residuals. The more complicated the model, the more choices there may be for $\hat{F}$.

Of particular, but not exclusive, interest to us here is the nonparametric case, where samples y^* drawn from $\tilde{F}$ will take the same values $y_1, \ldots, y_n$ as the data, but with variable frequencies $f_1^*, \ldots, f_n^*$. The right-hand side of (1) will often be evaluated by Monte Carlo simulation, which will generate R samples $y_r^* = (y_{r1}^*, \ldots, y_{rn}^*)$, $r = 1, \ldots, R$, and associated estimator values t_r^*. This notation extends in an obvious way to more complex data structures and semiparametric models. For general introductions see the books by Efron & Tibshirani (1993) and Davison & Hinkley (1997).

There are several situations in which standard bootstrap calculations and methods do not give reliable answers. In many situations corrective action is possible, by modifying the resampling scheme or by modifying some other aspect of the method. Here we briefly list problem situations and comment on some of the factors involved:

1. *Effect of data outliers.* Data outliers may exert considerable influence not only on the estimator T but also on the resampling properties of T^*. Depending upon the resampling model used, an outlier may occur with variable frequency in bootstrap samples, and the effect of this may be hard to anticipate even if T itself is not affected.

2. *Inconsistency of bootstrap method.* The combination of model, statistic and resampling scheme may be such that bootstrap results fail to approximate the required properties, no matter how large the sample size.

3. *Nonpivotality.* The key to success of the bootstrap is the accuracy of the substitution approximation (1). Often we have the choice of working quantity $U = u(Y, F)$, for example in confidence interval and significance test applications. The substitution principle is completely successful if U is a pivot, which is to say its distribution is the same for all F. This can be true exactly for parametric models, but only approximately for nonparametric models. It is important to know *how* approximately, and this can be done empirically.

4. *Incorrect resampling model.* For nonhomogeneous data it is necessary to correctly model the random variation, which in general means correctly identifying relevant strata according to which stratified resampling should be done. Examples of this occur with generalized linear models for count data, because of non-homogeneous residuals, and with temporal and spatial correlation, where for some methods the resampling scheme must respond appropriately to the range of dependence.

5. *Over-sensitivity to assumptions.* A bootstrap significance test often involves choosing a combination of statistic and method (e.g. transformation, studentization, acceleration and bias adjustment), as well as a resampling model. To the extent that there may be several similarly-reasonable resampling models, it would be important to check that the final result of the procedure not be overly-sensitive to which of these models might be chosen.

6. *Incorrect model-based calculation.* In semi-parametric problems, such as regression models with parametric structure for the mean and variance of response, approximate calculations for $\mathrm{var}(T^*)$ based on the variance model may lead to incorrect bootstrap calculations for studentized quantities if that variance model is wrong.

7. *Nonlinearity of statistic T.* Good properties of the bootstrap are often associated with smoothness of T as a function of $\hat{F}$, and in particular with the existence of an accurate linear approximation $T \approx \theta + n^{-1} \sum l(Y_j)$ or an appropriate generalization, in which $l(y)$ is the influence function of $t(\cdot)$ at (y, F).

8. *Discreteness of resample statistic T^*.* For very smooth statistics, the inherent discreteness of T^* (Hall, 1986) is not a problem, but for some rough statistics (such as the sample median or the sample maximum) the number of possible values of T^* presents difficulties which may require radical change in resampling scheme if reasonable results are to be obtained.

The focus of the present paper is on diagnostics, that is to say calculations that can be done with bootstrap output to diagnose the presence of one or more of the difficulties listed above. We focus particularly on items 1–4 in the above list, which we address in Sections 2–5 respectively. In Section 6 we include a brief illustration of item 5. The paper closes with a short discussion. For most of the numerical illustrations in this paper we have used the S-Plus library `boot`, which is accessible from `http://dmawww.epfl.ch/davison.mosaic/BMA/`.

Item 7 is discussed briefly in Section 3.10.2 of Davison & Hinkley (1997). Item 8 would usually be evident from plots of bootstrap output, and in many cases can be addressed using a smooth bootstrap as discussed in Section 3.4 of Davison & Hinkley (1997).

Although this paper emphasizes empirical diagnostics, it is important to stress the relevance of detailed theoretical work in this enterprise: this will be evident as we go along. It is clear that further theoretical work will be necessary before bootstrap methods become widely trustworthy additions to the statistical toolkit.

2 Data Outliers: Jackknife-after-bootstrap

One of the earliest accounts of diagnostics for bootstrap output was Efron's (1992) discussion paper, in which he introduced the "jackknife-after-bootstrap" plot. The idea of the plot is to examine the effect on bootstrap results (such as quantile or variance estimates) of deleting individual observations or cases. In the nonparametric case, rather than re-run bootstrap calculations for the dataset that results from single case-deletion, one can simply take the full-data bootstrap output and remove those resamples in which the case appears: this leaves roughly a fraction 0.632 of the resamples. The calculations are easily done if the table of resample frequencies f^*_{rj} (the number of times case j appears in the rth resample) is retained, or at least is easily reconstructed. In the parametric case a simple importance re-weighting can be used to augment the full-data bootstrap output to obtain case-deletion results; this is described at the end of this section.

Example: Survival data regression

As an illustration of the jackknife-after-bootstrap plot for quantile estimation, consider Figure 2. This relates to slope estimation for the data plotted in Figure 1, where there is fairly clearly one very large outlier, relative to any smooth curve, and possibly two moderate outliers relative to the linear regression that we fit to these data.

With x and y representing respectively dose and logarithm of survival rate, we fit the model $\mathrm{E}(Y \mid x) = \beta_0 + \beta_1 x$ to data $(x_1, y_1), \dots, (x_n, y_n)$ by least squares, obtaining estimates $\hat\beta_0$ and $\hat\beta_1$. There are now two main resampling methods:

(i) *Case Resampling:* simulate samples by randomly sampling from the empirical bivariate distribution of x, y,

$$(x^*_j, y^*_j) = (x_{U_j}, y_{U_j}), \quad j = 1, \dots, n,$$

with $U_1, \dots, U_n$ randomly sampled from $\{1, \dots, n\}$.

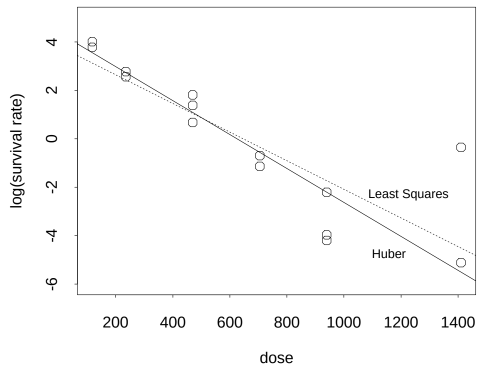

Figure 1: Survival percentages of rats at a succession of radiation doses (Davison & Hinkley, 1997, p. 308; Efron & Tibshirani, 1993, p. 116).

(ii) *Model-based Resampling:* simulate random samples from the fitted regression model, with the data design for x,

$$x_j^* = x_j, \quad y_j^* = \hat{\beta}_0 + \hat{\beta}_1 x_j + \varepsilon_j^*,$$

with errors ε_j^* randomly sampled from standardized residuals $e_j = (y_j - \hat{\beta}_0 - \hat{\beta}_1 x_j)/(1 - h_j)^{1/2}$, where h_j denotes the leverage for case j.

The top row of Figure 2 shows three different versions of the jackknife-after-bootstrap plot for quantiles of $\hat{\beta}_1^* - \hat{\beta}_1$ under case resampling, with $\hat{\beta}_1$ the least-squares slope estimate. The three plots correspond to three different estimates of the case influence values; see Appendix 1. The bottom row corresponds to model-based resampling. Case 13, the dominant outlier, shows up clearly in every plot for case resampling, aided by the fact that its influence value is extreme. But for model-based resampling, the outlier is somewhat obscured in two plots because of bad estimates of the influence value (due to the outlier, of course!). This is unfortunate, because it happens for the two most general methods of estimating influence values.

What the jackknife-after-bootstrap plot needs is some internal calibration, to indicate what outlier effects are significant, especially when the important points can be buried in the middle of the plot as above. A simple way to do this is to surround each dotted line on the plot, which is the full-data bootstrap quantile, with an uncertainty band for case-deletion quantiles that is computed using a robust estimate of standard deviation. This is illustrated in Figure 3, which corresponds to Figure 2. The width of each band here is 1.645 times the standard deviation based on the interquartile range of the case-deletion quantiles. There are other ways to calibrate the plot, such as using box-plots of quantiles ("jack-in-the-box" plots), or simulation envelopes for the quantiles, but the bands are both clear and simple.

For this example, two main options for the data analysis are (i) omission of case 13 (the major outlier) and re-application of the least squares fit, and (ii) replacement of the least squares fit by a robust fit. Under the latter option it is important to consider what effect the outlier might still

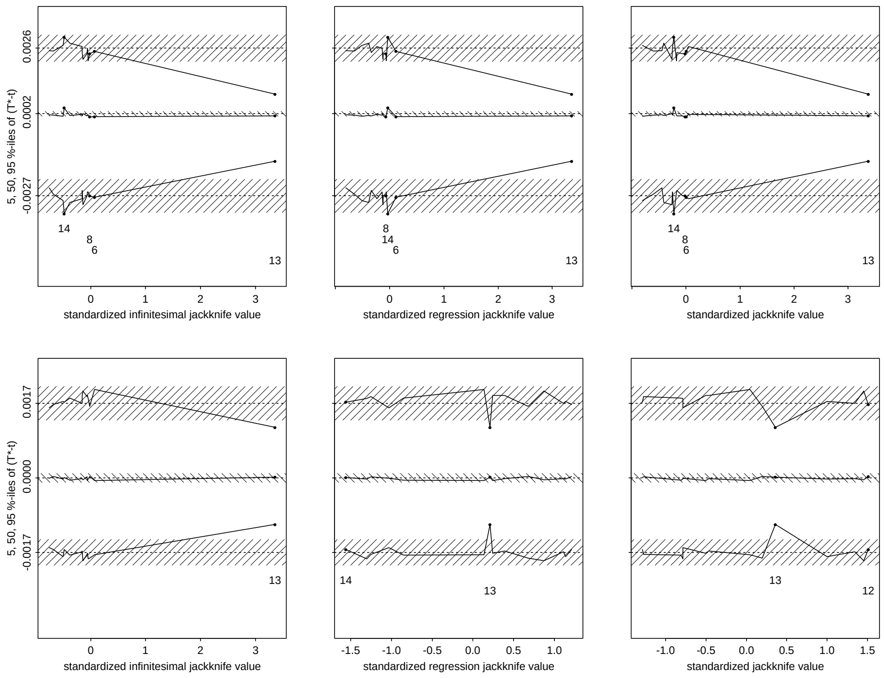

Figure 2: Jackknife-after-bootstrap plot for survival data least-squares regression slope, for data shown in Figure 1. $R = 999$. Top row: case resampling. Bottom row: model-based resampling.

have on the bootstrap analysis, even if the fit itself is not affected. As Singh (1998) points out, the breakdown point for bootstrap analysis may be quite different to that for the data fit itself. For example, suppose that case j is an outlier, and that the particular robust estimate t is not sensitive to this one outlier. Bootstrap samples under case resampling will contain case j with frequencies that usually range from 0 to 3, and if the estimator T is vulnerable to two or more outliers, then some values of t^* will be affected – thereby contaminating moment estimates, and possibly also quantile estimates. The effects may differ appreciably between case and model-based resampling, as happens for the data in this example.

The left panel of Figure 4 shows a jackknife-after-bootstrap plot for the Huber Proposal 2 estimate of slope as calculated by the S-Plus function `rlm` of Venables & Ripley (1997), when case resampling is used. It is unclear how to do model-based resampling in this context. Table 1 contains 0.05 and 0.95 quantiles of $\hat{\beta}_1^* - \hat{\beta}_1$ under case resampling, for Huber estimation and for least squares estimation when case 13 is omitted, which the Huber estimate effectively mirrors. Notice the large differences. Figure 4 shows very clearly that case 13 has a large effect on bootstrap analysis of the robust estimate, despite having virtually no influence on the numerical value of the estimate itself.

This difficulty with resampling robust estimates in the presence of outliers may be circumvented by use of weighted resampling, in which case weights of the robust fit are used as resampling weights. For the present example, the final weights in the iterative weighted least squares calculation of `rlm`

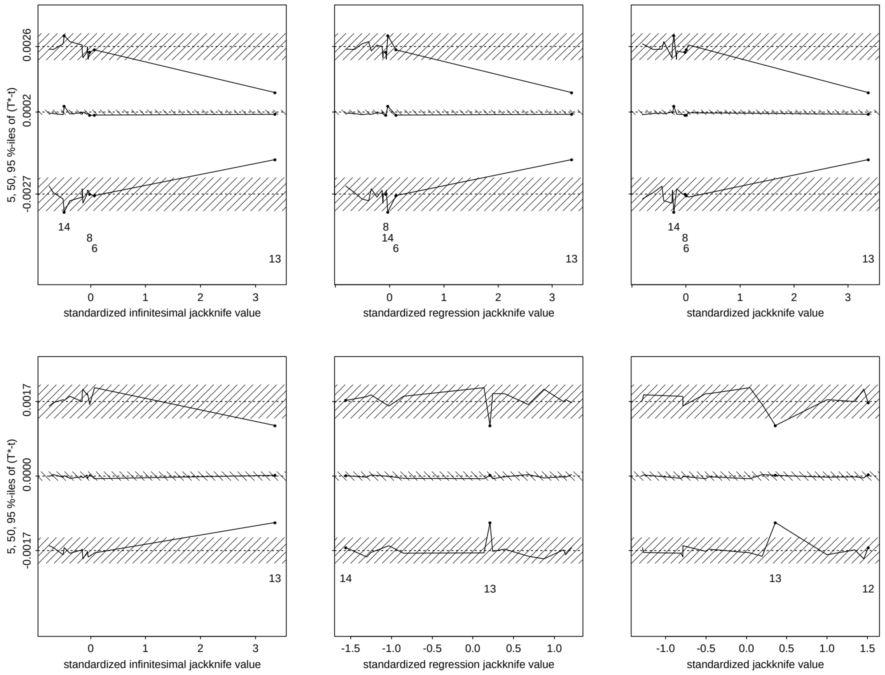

Figure 3: Jackknife-after-bootstrap plot with calibration bands for survival data least-squares regression slope. Top row: case resampling, bottom row: model-based resampling. $R = 999$.

Table 1: Quantiles of $\hat{\beta}_1^* - \hat{\beta}_1$ for survival data, case resampling $R = 999$.

	5% quantile	95% quantile
Huber estimate	-0.0021	0.0032
least squares estimate	-0.0018	0.0008

are

$$1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 0.40, 0.46, 0.16, 1.00$$

The right panel of Figure 4 contains the jackknife-after-bootstrap plot for weighted resampling using the above weights; we have matched the vertical scales of the two panels to facilitate comparison. Clearly case 13 no longer has damaging effects on the resampling.

The difficulty with this circumvention is that it is not available for all robust estimators: for example, it cannot be used with least trimmed squares estimators, although then the bootstrap analysis appears to be stable and reliable.

□

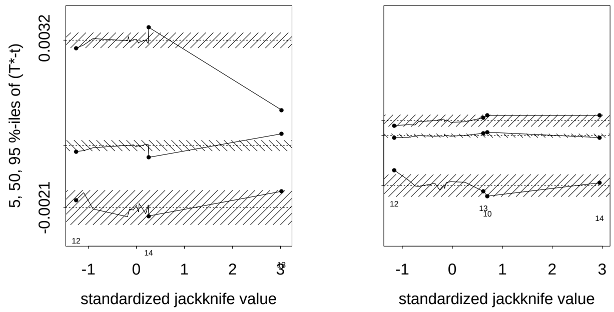

Figure 4: Jackknife-after-bootstrap plots for 5, 50 and 95% quantiles of Huber slope estimate. Case resampling, $R = 999$. Left panel: equal weights; right panel: robust weights. Both vertical scales are the same.

Parametric case

The case-deletion calculations for parametric bootstraps are necessarily different. They make use of the following "importance weighting" formula

$$\mathrm{E}\{h(Y^*) \mid \hat{\psi}_{-j}\} = \mathrm{E}\left\{ h(Y^*) \frac{f(Y^* \mid \hat{\psi}_{-j})}{f(Y^* \mid \hat{\psi})} \,\middle|\, \hat{\psi} \right\} \equiv \mathrm{E}\{h(Y^*) w_j(Y^*) \mid \hat{\psi}\}, \tag{2}$$

where $\hat{\psi}$ and $\hat{\psi}_{-j}$ are typically maximum likelihood estimators, respectively from the full data and from the data with case j deleted. For the cumulative distribution of T^*, (2) is used with $h(y^*)$ the appropriate indicator function with argument t^*, which leads to the following algorithm for case-deletion quantiles (Davison & Hinkley, 1997, Section 9.4):

- for each resample y_r^* from $\hat{F} \equiv F_{\hat{\psi}}$, calculate the set of weights $w_{r,j}^* = w_j(y_r^*)$, for $j = 1, \ldots, n$, according to (2),
- rearrange resample subscripts so that $t_1^* \leq t_2^* \leq \cdots \leq t_R^*$,
- for any $0 < p < 1$ of interest, define $s(j,p) = \min\left\{ S : \sum_{r=1}^{S} w_{r,j}^* \geq p\left(\frac{R+1}{R}\right) \sum_{r=1}^{R} w_{r,j}^* \right\}$, for $j = 1, \ldots, n$,
- calculate the p quantile of $T^* - t$ under $\hat{F}_{-j}$ as $t_{s(j,p)}^* - t$.

Note that if $\hat{\psi}_{-j} = \hat{\psi}$, then $t_{s(j,p)}^* = t_{(R+1)p}^*$, the full-data quantile. The weights $w_{r,j}^*$ are implicitly normalized to have average one, to match their theoretical expectation: other adjustments are possible.

3 Inconsistency of bootstrap method

Cases of bootstrap inconsistency include the sample maximum (Politis & Romano, 1994, after Bickel & Freedman, 1981); the Hodges superefficient location estimate and the Stein estimate for a multivariate mean (Beran, 1997); and kernel estimates for densities and for regression curves (Härdle & Bowman, 1988, Härdle & Marron, 1991). In fact there are two types of inconsistency: non-existence of the quantity being estimated by the bootstrap – a sort of "reverse inconsistency" – and the failure of the bootstrap to agree with an existent quantity even in the limit as the sample size (or equivalent) grows to infinity, i.e. a failure of the substitution approximation (1). The latter seems more important and we discuss this first.

Absolute inconsistency

Inconsistency, where the bootstrap will typically fail to give a distribution estimate that is consistent for the true distribution, requires rather subtle analysis. One major theoretical contribution, by Beran (1997), concerns the relation between superefficiency and inconsistency. A rough summary of Beran's findings is that if T is a maximum likelihood estimator (literally, in the parametric case, or $T = t(\hat{F})$ with $\hat{F}$ the nonparametric maximum likelihood estimator of F in the nonparametric case), and if $\tilde{T}$ is a superefficient estimator, with the superefficiency restricted to the parameter subset Θ_S, then the distribution of $n^{1/2}(\tilde{T}^* - \tilde{T})$ will not be a consistent estimator of the distribution of $n^{1/2}(\tilde{T} - \theta)$ for $\theta \in \Theta_S$ if resampling is done under the model with parameter value T or $\tilde{T}$.

Theorem 2.2 of Beran's paper suggests a diagnostic plot to reveal the lack of consistency. Define $A^* = \tilde{T}^* - T^*$ and $D^* = T^* - t$. These quantities should be approximately independent if the bootstrap distribution of $\tilde{T}^* - \tilde{T}$ is a consistent approximation for the distribution of $\tilde{T} - \theta$. The diagnostic plot is therefore a plot of values of $a^* = \tilde{t}^* - t^*$ versus $d^* = t^* - t$. Note that these quantities might be rescaled or transformed, and if they are vectors then suitable scalar functions of them will need to be chosen for plotting.

Beran illustrates this method with the superefficient Hodges estimator for a normal scalar mean. An example of more practical interest is the Stein estimator for a multivariate mean, which serves as a prototype for many nonparametric smoothers.

Example: Stein estimator of Normal means

Suppose that we observe $y = (y_1, \ldots, y_n)$ such that the y_j are independent $N(\theta_j, 1)$ measurements for $j = 1, \ldots, n$, with $n > 3$. We want to calculate a confidence set for the vector of means $\theta = (\theta_1, \ldots, \theta_n)$. The maximum likelihood estimator for θ is just $t = y$. An alternative, and in some cases much better, estimator is the Stein estimator $\tilde{t}$ with jth component

$$\tilde{t}_j = \bar{y} + \left\{1 - \frac{(n-3)}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\right\}(y_j - \bar{y}). \tag{3}$$

This estimator is superefficient in terms of total mean squared error, everywhere dominating the maximum likelihood estimator $t = y$: the strongest superefficiency is on the set $\Theta_S = \{\theta : \theta_1 = \cdots = \theta_n\}$, where $\mathrm{E}\{(\tilde{T} - \theta)^T(\tilde{T} - \theta)\} = 3 < n = \mathrm{E}\{(T - \theta)^T(T - \theta)\}$.

Whichever estimator is used, it is natural to form a confidence set for θ as a sphere based on the sum of squared errors of the estimator. In the case of the maximum likelihood estimator, bootstrap analysis for $u = \sum_j (t_j - \theta_j)^2$ with resampling from the $N(t, I)$ distribution will be consistent –

Table 2: Simulated data sets ($n = 10$) for Stein estimation example. First three rows are samples for the null case $\theta = 0$. Last three rows are samples for the case $\theta = (1, \ldots, 10)$.

$\theta = 0$	-1.02	-0.98	-0.42	-0.18	-0.18	0.25	0.48	0.89	1.76	1.90
$\theta = 0$	-1.20	-0.75	-0.44	-0.21	-0.21	0.09	0.28	0.35	0.64	1.39
$\theta = 0$	-1.51	-1.12	-0.88	-0.67	-0.39	-0.09	0.03	0.28	0.44	1.02
$\theta = (1, \ldots, 10)$	1.22	2.74	3.18	3.48	5.10	5.61	6.54	6.96	7.87	11.87
$\theta = (1, \ldots, 10)$	0.53	1.98	3.00	3.52	5.19	6.06	6.39	7.33	9.03	9.16
$\theta = (1, \ldots, 10)$	0.52	1.92	2.30	2.57	5.58	5.86	6.48	8.98	9.54	10.56

in fact U is a pivot having the χ^2_n distribution for all θ. But when the Stein estimate $\tilde{t}$ is used, resampling can be shown (Beran, 1995) to give inconsistent results under resampling from the $N(\tilde{t}, I)$ distribution (as the substitution principle would require) or from the $N(t, I)$ model, unless $\sum_j (\theta_j - \bar{\theta})^2$ is very large. Beran does discuss (partially sucessful) methods for dealing with the difficulty. Here we focus on the ability to diagnose the inconsistencies by empirical analysis.

To examine the use of Beran's diagnostic plot, we consider the raw difference $\tilde{T} - \theta$ and its bootstrap analog $\tilde{T}^* - \tilde{t}$. For the diagnostic plot we take

$$a^* = \sum_{j=1}^{n} (\tilde{t}_j^* - t_j^*)^2, \quad d^* = \sum_{j=1}^{n} (t_j^* - t_j)^2; \tag{4}$$

logarithms may be used to better spread the scatterplot for small n. Figure 5 illustrates the case of $n = 10$ null means, when the data are independently sampled from the $N(0, 1)$ distribution, i.e. all means are equal and $c = 0$. The figure shows plots for the three different data sets in the top half of Table 2, and for each is shown the results from resampling $R = 100$ times with the maximum likelihood fitted model (left), the Stein fitted model (center), and another model to be described in Section 6. The bottom row is also for the third data set, but with $R = 1000$ resamples rather than $R = 100$. The dependence is stronger in some cases than others, but is evident for most data sets. Clearly $R = 100$ is sufficient for diagnostic purposes.

Figure 6 shows corresponding plots for the three non-null data sets in the bottom half of Table 2: for these $\theta = (1, 2, \cdots, 10)$. The pattern in these plots is very different: here $\sum_j (\theta_j - \bar{\theta})^2$ is very large, and the expected approximate independence is clear.

$\square$

Similar behaviour to that of the preceding example may occur with kernel nonparametric regression, where estimates are implicitly of shrinkage form. It is well-known (Härdle & Marron, 1991) that successful model-based resampling requires both fitted curve and residuals to correspond to bandwidths different from those used for the data estimate itself. It would be interesting to see corresponding diagnostic analysis for examples of such nonparametric regression. Similar behaviour is also observed with wavelet estimation of regression curves based on thresholded coefficients, which

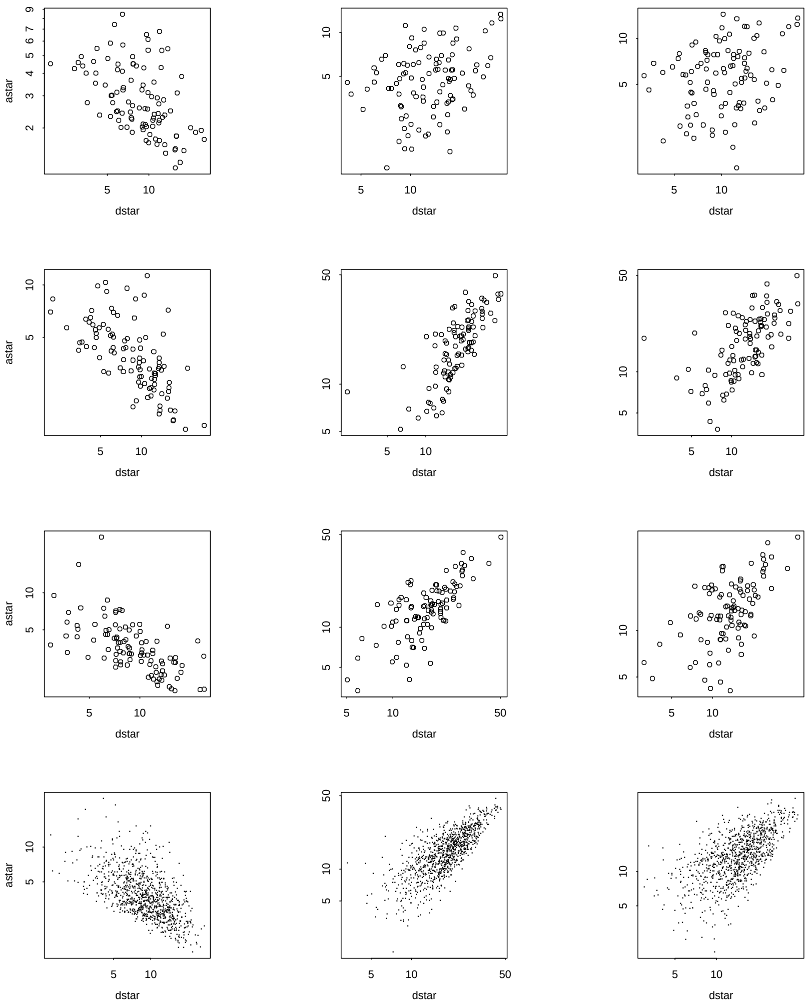

Figure 5: Diagnostic plot for failure of standard parametric bootstrap for Stein estimator, $n = 10, \theta = 0$. Statistics defined in (4). Maximum likelihood resampling model (left column), Stein estimate resampling model (center column), Beran estimate resampling model (right column). Rows 1–3 correspond to three data sets with $R = 100$, and row 4 corresponds to the third data set but with $R = 1000$.

we discuss in Section 5. The difficulties in the nonparametric context are in choosing an appropriate analog of the maximum likelihood estimate against which to contrast the estimate of interest (the analog of the superefficient estimate), and in completing the descriptions of the possible resampling models by appropriate choices of error component.

Relative inconsistency

A more subtle type of inconsistency is evident in the work of Lee & Young (1995) on nonparametric bootstrap inference for the mean when data are sampled from a lognormal distribution. Their

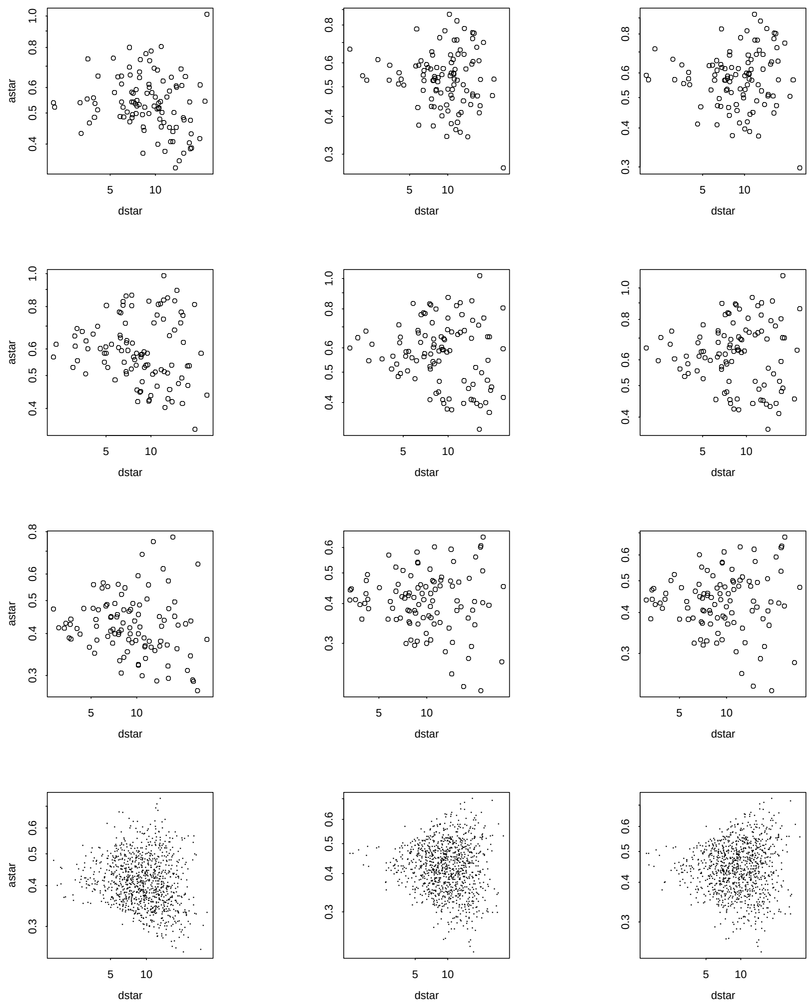

Figure 6: Diagnostic plot for failure of standard parametric bootstrap for Stein estimator, $n = 10, \theta = (1, \cdots, 10)$. Statistics defined in (4). Maximum likelihood resampling model (left column), Stein estimate resampling model (center column), Beran estimate resampling model (right column). Rows 1–3 correspond to three data sets with $R = 100$, and row 4 corresponds to the third data set but with $R = 1000$.

expansions for coverage of the main bootstrap confidence interval methods suggest strange behavior, which is dramatically confirmed in simulation studies. For example, one of the best (second-order correct) nonparametric confidence interval methods has the theoretical property that the nominal 90% interval $I_{0.90}(\tilde{F})$ has coverage

$$\Pr\{\theta \in I_{0.90}(\tilde{F})\} \approx 0.90 - 2.5 \times 10^{20} n^{-2},$$

which is wildly inaccurate as an approximation. Correspondingly, for finite sample sizes $n = 20, 35, 100$ the empirical coverages are 53%, 60%, 70% respectively. Jackknife-after-bootstrap plots as described in Section 2 typically suggest difficulty with bootstrapping the estimated mean (and variance) for lognormal data, for both parametric and nonparametric resampling.

Wood (1998) relates this inaccuracy of consistent methods to subexponentiality of the data distribution, a characteristic possessed by the lognormal distribution. The essence of the problem is as follows. Consistency in the usual absolute sense provides that if $\Pr(T-\theta \geq u_{1-\alpha}) = \alpha$, with α fixed, then the bootstrap estimate $\hat{u}_{1-\alpha} \to u_{1-\alpha}$ as $n\to\infty$. But more relevant to actual coverage of finite sample upper confidence limits for θ may be the ratio

$$\rho(\alpha_n; \hat{F}, F) = \frac{\Pr(T^* - t \geq u_{1-\alpha_n} \mid \hat{F})}{\alpha_n} \tag{5}$$

and its limiting behavior when $\alpha_n \to 0$ as $n\to\infty$: it is believed that finite sample accuracy requires that this ratio tends to 1. Wood obtains theoretical results concerning the value α_n at which the convergence of the ratio to 1 breaks down. There are corresponding results for the left tail, but for positive data these are not as relevant.

The difficulty as far as an empirical diagnostic is concerned is that $\rho(\alpha_n; \hat{F}, F)$ fails to tend to one because the normal approximation for the parametric denominator of (5) fails, but the numerator normal approximation does not. We have not been able to devise a wholly empirical nonparametric analog of $\rho(\alpha_n; \hat{F}, F)$ with the same behavior.

Reverse inconsistency

In a 'typical' simple problem one expects that $U_n = n^{1/2}(T_n - \theta)$ has distribution G_n with a proper limit as $n\to\infty$ and such that $U_n^* = n^{1/2}(T_n^* - t_n)$ has distribution $\hat{G}_n$ converging to the same limit as $n\to\infty$. Most often the limit is a Normal distribution. This being the case, one would expect that if we took a resample of size $m < n$ from the sample of size n, and calculated $T_{n,m}^*$ from that resample, then $U_{n,m}^* = m^{1/2}(T_{n,m}^* - t_n)$ would have distribution $\hat{G}_{n,m}$ close to $\hat{G}_n$. So to check whether or not the standardization $n^{1/2}$ is correct, it would seem appropriate to plot empirical quantiles of U_n^* against corresponding quantiles of $U_{n,m}^*$ for several m smaller than n: a linear trend with slope near one should be seen if all is well. Problems to which this might apply include extreme value statistics, nonparametric density and regression estimates, and averages of heavy-tailed distributions. Other examples including the mode are discussed by Léger and MacGibbon (1998).

A more constructive development by Bertail, Politis & Romano (1999) is an empirical subsampling method for determining the rate of convergence of $T_n - \theta$. Suppose that for some $\alpha < 0$, $T_n - \theta = O_p(n^{\alpha})$. Then for any probability p the pth quantile of $T_n - \theta$ is $O_p(n^{\alpha/2})$. This suggests that, at least for large n, a reasonable estimate for α can be obtained as twice the slope from a plot of the logarithm of the pth quantile of $T_{n,m}^* - t$ versus $\log m$.

Example: Mean and median for long-tailed distributions

To assess the potential of this approach as a diagnostic tool, we applied it to samples of size $n = 100$ from the t_ν distribution. (For small ν enormous observations can appear, so those greater than 40 in absolute value have been deleted: thus, although they possess moments of all orders, the data are

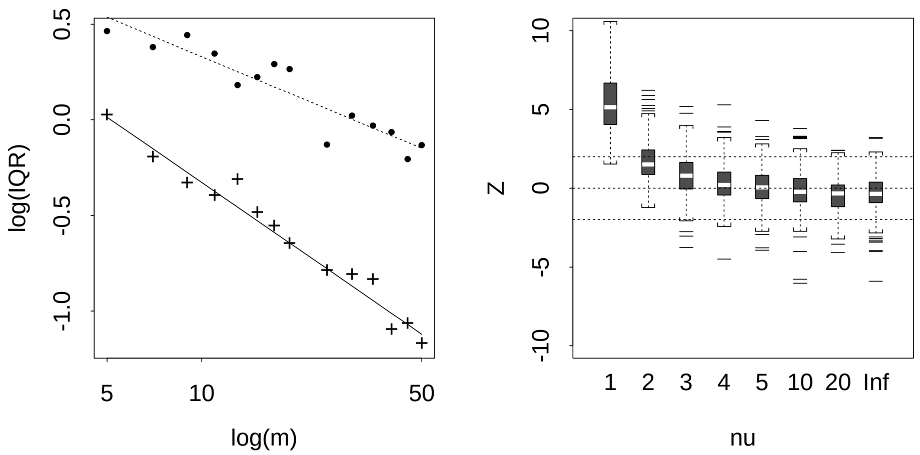

Figure 7: Subsampling diagnostic for bootstrap failure of the average. Left: $\log IQR$ for averages ($\bullet$) and medians ($+$) of 199 subsamples of size m from a sample of size 100 from the Cauchy distribution. Right panel: Empirical distribution of a diagnostic for failure of the bootstrap, for samples from the t distribution with ν degrees of freedom; horizontal lines are at $0, \pm 2$.

essentially indistinguishable from t_ν samples.) We calculated the average and median of $R = 199$ subsamples of size $m = 3, 5, \ldots, 19$ taken without replacement from each sample. For each m we use the interquartile range IQR of the 199 subsamples as a robust summary of their spread. The median has variance $O(m^{-1})$ for all m, but the average has finite variance $O(m^{-1})$ for t_ν samples only if $\nu \geq 3$. Hence comparison of the slopes of graphs of $\log IQR_{\rm ave}$ and $\log IQR_{\rm med}$ against $\log m$ should be informative about whether it is safe to bootstrap the average; a natural way to assess this is the t statistic for slope, $Z_{\rm sub}$, when $\log(IQR_{\rm ave}/IQR_{\rm med})$ is regressed on $\log m$.

The left panel of Figure 7 shows how this works for a sample generated with $\nu = 1$; we took $m = 25, 30, \ldots, 50$ in addition to the values mentioned above. The slopes and standard errors of the fitted lines for the average and median are respectively -0.3 (0.03) and -0.49 (0.03), and $Z_{\rm sub} = 4.63$, giving apparently strong evidence of different rates of convergence. However the evidence would be weaker if only values $m = 25, \ldots, 50$ had been used; for this subset $Z_{\rm sub} = 2.99$. As m increases the dominant effect becomes that of sampling from a finite population, which of course has moments of all orders, and for some samples there is clear downward curvature in the subsampling results for the average, with slopes for the average and median essentially equal for large m. This suggests that m needs to be very small compared to the sample size, and hence that very large samples are needed for subsampling to work well. Our experience with this method is not as good as that of Bertail, Politis & Romano (1999), who considered samples with $n = 100$, 1000 and 10,000.

The right panel shows 200 replicates of $Z_{\rm sub}$ for various ν. There is a noticeable upward shift when $\nu \leq 2$, though even with normal samples ($\nu = \infty$), the variance of $Z_{\rm sub}$ is larger than one.

About 60% of samples have $Z_{\text{sub}} > 2$ when $\nu \leq 2$, dropping to about 6% for larger ν. Thus a rule-of-thumb that the average should not be bootstrapped if $Z_{\text{sub}} > 2$ would be conservative, but not badly so. The usual preliminary to location analysis would be a normal scores plot of the sample. If this showed long tails it would be wise to replace the average with a robust estimator, independent of the results of a subsampling analysis. Subsampling is likely to be most valuable for more complex situations than location estimation, but using independent data.

$\square$

Subsampling has been proposed for use with time series, using blocks of m consecutive observations. There are then just $n - m + 1$ subsamples, and the corresponding values of the statistic are highly correlated. Hence quantities based on the subsampled statistics, such as the IQR, tend to be highly variable, making it very difficult to extract reliable information from plots such as the left panel of Figure 7. Though limited, our experience suggests that huge samples of stationary data would be needed for this form of subsampling to work well.

4 Pivotness

The key to success of the bootstrap is the accuracy of the substitution principle (1). Usually we have the choice of working quantity $U = u(Y, F)$, for example the estimation error on a transformed scale $U = n^{1/2}\{h(T) - h(\theta)\}$ or the studentized form $Z = (T - \theta)/V^{1/2}$ with V an appropriate variance estimate. Suppose that $G_n(u)$ is the CDF of U. We know that in the theoretical world $\hat{G}_n(u) = G_n(u) + O_p(n^{-1/2})$, at least for regular problems where the bootstrap is consistent. But this is not strong enough to make bootstrap methods accurate: we need at least $\hat{G}_n(u) = G_n(u) + O_p(n^{-1})$, preferably $\hat{G}_n(u) = G_n(u)\{1 + O_p(n^{-1})\}$, and ideally $\hat{G}_n(u) = G_n(u)$ for exact pivotness. The choice $U = Z$ will generally give improved accuracy, and sometimes $U = n^{1/2}\{h(T) - h(\theta)\}$ will do the same for some carefully-chosen transformation $h(\cdot)$. But quite aside from what may be true asymptotically, we need to know just how close we are to the ideal before we can rely on quantiles of $\hat{G}_n$.

In a parametric setting the notion of pivot is firm: the quantity $U = u(Y, F)$ would have the same distribution for all parameter values. This could be checked empirically by resampling from models F_θ with θ varying over a wide range, including the estimated value t, and then plotting empirical characteristics such as mean, variance and quantiles against θ.

In the nonparametric case almost nothing interesting will be an exact pivot, but we can think informally of approximate pivotness as the property that $U = u(Y, F)$ will have approximately the same distribution for all F near the EDF $\tilde{F}$ (and its generalizations). This we can assess by examining the distribution of $U = u(Y, F)$ for the special distributions $\tilde{F}^*$ obtained by resampling, on the grounds that these distributions must encompass the same neighborhood that almost certainly includes F itself. Taken literally, this idea involves us in double bootstrapping, but as we shall see this can often be avoided. An alternative approach, which in practice turns out to be less successful, is to embed $\tilde{F}$ in an empirical exponential family by exponential tilting (Davison & Hinkley, 1997, Section 9.4).

Consider first the problem of checking the stability of the bias and variance of T. The bootstrap estimates of these quantities, obtained by resampling R times from $\tilde{F}$, are

$$b = \bar{t}^* - t, \quad v = (R-1)^{-1}\sum_{r=1}^{R}(t_r^* - \bar{t}^*)^2.$$

These correspond to the parameter value $\theta = t$ of $\tilde{F}$. To calculate the analogous values at parameter value t_r^* we resample from $\tilde{F}_r^*$, the EDF of the rth bootstrap sample $y_{r1}^*, \ldots, y_{rn}^*$. If we draw M samples from $\tilde{F}_r^*$, then estimates of the bias and variance are

$$b_r^* = \bar{t}_r^{**} - t_r^* = M^{-1}\sum_{m=1}^{M} t_{rm}^{**} - t_r^*, \quad v_r^* = (M-1)^{-1}\sum_{m=1}^{M}(t_{rm}^{**} - \bar{t}_r^{**})^2.$$

We would now plot these against values of t_r^* as a surrogate for plotting bias and variance of T against θ. Examples of such plots are shown in Tibshirani (1988), Sections 3.9 and 9.4 of Davison & Hinkley (1997) and Ventura (1997, 1998).

There are two problems with the simulation calculations as just outlined. One is that potentially the total number of simulations $(RM + R)$ is very large. The other is that the resulting graph is typically noisy because of the discreteness of the $\tilde{F}_r^*$ and their lack of smoothness with respect to t_r^*. We deal with these difficulties in two ways, first by smoothing the $\tilde{F}_r^*$ and secondly by using a recycling strategy that involves sampling from many fewer than all R bootstrap EDFs.

The smoothed version of $\tilde{F}^*$ that corresponds to parameter value θ has probabilities on the data values $y_1, \ldots, y_n$ defined by (Davison, Hinkley & Worton, 1995)

$$p_{\theta,j}^* \propto \frac{1}{R}\sum_{r=1}^{R} f_{rj}^* w\left(\frac{t_r^* - \theta}{\varepsilon}\right), \quad j = 1, \ldots, n, \tag{6}$$

where f_{rj}^* is the frequency with which y_j appears in the rth bootstrap resample, $w(\cdot)$ is a symmetric probability density with mean zero and unit variance, and ε is an appropriate smoothing parameter; the constant of proportionality in (6) is chosen so that $\sum_{j=1}^n p_{\theta,j}^* = 1$. We denote the corresponding cumulative distribution function by F_θ^*; the corresponding parameter value $t(F_\theta^*)$ does not exactly equal θ, but for small values of ε – we use ε in the range 0.2–$1.0v^{1/2}$ – there will be negligible difference. We use the standard normal density for $w(\cdot)$.

In principle we could use (6) with $\theta = t_r^*$ to replace $\tilde{F}_r^*$ for every r, but in practice we take, say, 20 to 100 values of θ equally spaced over the central 95% range of the t_r^* values.

Having fixed on the distributions F_θ^* that replace the $\tilde{F}_r^*$s, we now introduce recycling to avoid the separate resampling from all F_θ^*s. The recycling idea is to make repeated use of the importance sampling identity

$$\kappa(H) \equiv \mathrm{E}\{k(X) \mid H\} = \mathrm{E}\left\{ k(X)\frac{dH(X)}{dG(X)} \middle| \, G \right\}, \tag{7}$$

with a fixed G that assigns probabilities $g_1, \ldots, g_n$ respectively to the data values $y_1, \ldots, y_n$ and H ranging over all F_θ^* of interest, $X = (Y_1^*, \ldots, Y_n^*)$. The expectations in (7) are replaced by averages over Q samples drawn from G, so that for all values of θ considered, $\mathrm{E}\{k(Y^*) \mid F^*_\theta\}$ is approximated by

$$\hat{\kappa}(\theta) = Q^{-1} \sum_{q=1}^{Q} k(y^*_q) w^*_{\theta,q}, \tag{8}$$

where

$$w^*_{\theta,q} = \prod_{j=1}^{n} \frac{d\tilde{F}^*_\theta(y^*_{qj})}{dG(y^*_{gj})} = \prod_{j=1}^{n} \left(\frac{p^*_{\theta,j}}{g_j} \right)^{f^{**}_{qj}}, \tag{9}$$

with f^{**}_{qj} equal to the frequency of data value y_j in the qth resample from G. The simplest choice for G is $\tilde{F}$, for which the resamples y^{**}_q are just ordinary first-level bootstrap resamples, but more effective choices involve mixtures of $\tilde{F}$ with a small number of the $\tilde{F}^*_r$ (Ventura, 1997, 1998).

The use of smoothed estimates F^*_θ in place of the original resample EDFs $\tilde{F}^*_r$ has a dramatic effect on this recycling algorithm, which otherwise would be extremely unstable (Ventura, 1997, 1998). For example, if $n = 10$ and we use unsmoothed distribution $\tilde{F}^*_r$ with $G = \tilde{F}$, then the recycling weights $w_{r,q}$ analogous to (9) would nearly all (96%) be zero while having mean 1.

Yet further improvement can be made to the simulation by using the identity $\mathrm{E}(W) = 1$ as a control, as discussed by Hesterberg (1995) and Ventura (1997, 1998). Simplest is the ratio method, whereby the simple basic recycling estimate (8) is replaced by

$$\hat{\kappa}_{\mathrm{ratio}}(\theta) = \frac{\sum_{q=1}^{Q} k(y^*_q) w^*_{\theta,q}}{\sum_{q=1}^{Q} w^*_{\theta,q}}. \tag{10}$$

The regression method makes a covariate adjustment to give

$$\hat{\kappa}_{\mathrm{reg}}(\theta) = Q^{-1} \sum_{q=1}^{Q} k(y^*_q) w^*_{\theta,q} - c(\bar{w}^* - 1),$$

with c the coefficient of linear regression of $k(Y)W$ on W.

These ideas have been implemented in an S-Plus function `pivot.plot`, as an adjunct to our S-Plus library `boot`, which plots estimated bias, standard error and quantiles of $T^* - t$ versus t for a range of t values. Details of the implementation are summarized in Appendix 2.

One simple application of a primitive form of these ideas with a ratio estimate is discussed in Sections 3.9.2, 3.10.2, and 9.4.4 of Davison & Hinkley (1997). Here we examine a more complicated example.

Example: Insurance data

The left panel of Figure 8 shows an exponential order statistics plot of extreme values for some insurance claim data. The vertical axis shows ordered values of $y_j = x_j - u$, where the x_j are the 254 insurance claims greater than a threshold of $u = 5$ units in the data on Danish fires shown in Figure 6.2.11 of Embrechts, Kluppelberg & Mikosch (1997). The statistical problem here is inference for the right tail of the data, corresponding to the largest insurance claims. One approach

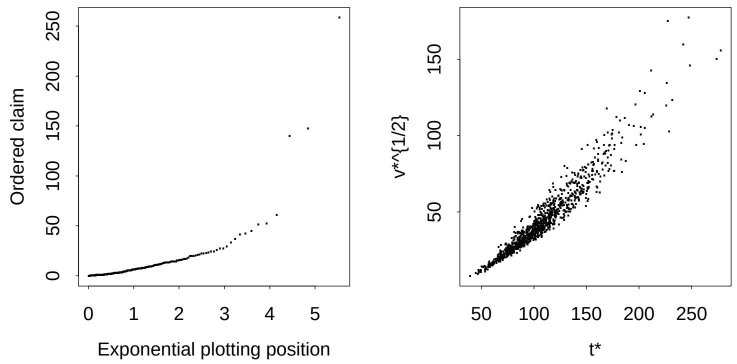

Figure 8: Insurance claim data. Left panel: plot of ordered data against exponential order statistics. Right panel: plot of $v^{*1/2}$ against t^* for 0.99 quantile estimate $t = \hat{\xi}_{0.99}$ of claim distribution, using $R = 999$ bootstrap replicates.

to analysis of such data via an extreme-value paradigm is a maximum likelihood fit to the y's of the generalized Pareto distribution

$$F(y) = \begin{cases} 1 - \exp(-y/\sigma), & \kappa = 0, \\ 1 - (1 + \kappa y/\sigma)_+^{-1/\kappa}, & \text{otherwise}, \end{cases} \quad 0 < \sigma, -\infty < \kappa < \infty. \tag{11}$$

The p quantile ξ_p of F is then estimated using $\xi_p = \sigma\{(1-p)^{-\kappa} - 1\}/\kappa$, with parameters κ and σ replaced by maximum likelihood estimates. For illustration we take $p = 0.99$, which gives $\hat{\xi}_p = 104.5$, using $\hat{\sigma} = 3.81$ and $\hat{\kappa} = 0.63$. A confidence interval for ξ_p can be obtained using a profile likelihood, but this is based on the model, and as (11) is a high-threshold approximation to the distribution of the y's, one might question whether $u = 5$ is high enough for the model to fit well. A more model-robust analysis uses the estimator $t = \hat{\xi}_p$ of the underlying quantile, but assesses its uncertainty via the nonparametric bootstrap.

An estimated variance v for $t = \hat{\xi}_p$ is obtained by the nonparametric delta method, $v = n^{-2}\sum l_j^2$, with l_j the exact empirical influence value for case j determined by the maximum likelihood estimating equations for κ and σ and the definition of ξ_p; see Problem 2.12 of Davison & Hinkley (1997).

Confidence intervals for ξ_p can be based on the studentized statistic $Z^* = (T^* - t)/V^{*1/2}$, where T^* and V^* are the versions of t and v calculated on bootstrap samples from $y_1, \ldots, y_n$. One might also consider transforming $t = \hat{\xi}_p$ before studentizing. In this data set Z^* has a very skewed distribution, however, and it is natural to ask if it is close to pivotal, particularly as the plot of $v^{*1/2}$ versus t^* in the right panel of Figure 8 shows such a strong relationship. Here we consider the somewhat simpler question of whether or not $T^* - t$ is pivotal; and if not, does a logarithmic transformation cure the problem?

Figure 9 shows the results from applying `pivot.plot` to bootstrap output for t^* obtained with $R = 1000$, $S = 500$ additional samples in each tail, and $\varepsilon = 0.02SE$ in the frequency smoothing that precedes calculation of the importance sampling weights. The plot suggests strongly that $T^* - t$ is not close to pivotal. Figure 10 shows the corresponding plot for $\log t^*$, which suggests that the logarithmic transformation stabilizes the distribution quite well. Both plots include "error bands" designed to indicate the level of simulation error (due to variability of bootstrap samples) : the bands are drawn at ± 1 standard errors of the central bootstrap value. Concrete illustrations of the accuracies of these plots are provided in Figures 11 and 12: in both, the right panel shows ten independent replicates of the quantile part of `pivot.plot`, while the left panel show five independent replicates of the corresponding full double bootstrap version, this using $R = 500$ and $M = 1000$ second-level resamples for each first-level resample.

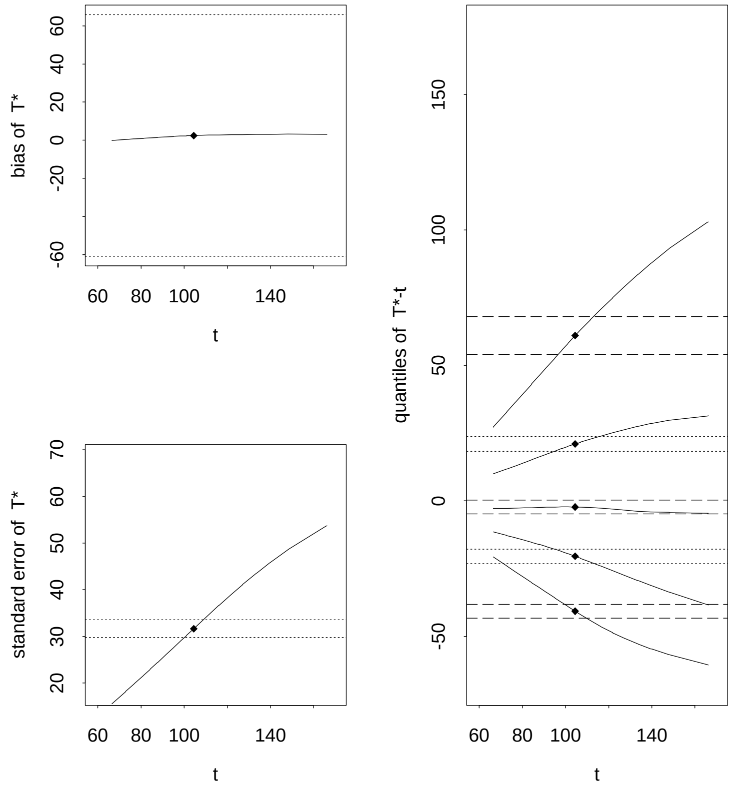

Figure 9: Insurance claim data. Result of `pivot.plot` for estimating the dependence on $\xi_{0.99}$ of bias (top left panel) and standard error (bottom left panel) of $\hat{\xi}_{0.99}$, and of 5, 25, 50, 75, 95 % quantiles of $\hat{\xi}_{0.99} - \xi_{0.99}$.

As one can tell from Figures 9 and 10, it is important to look at the vertical scales (and the error bands) on these plots rather carefully, so as to not read significance into differences that are inconsequential.

$\square$

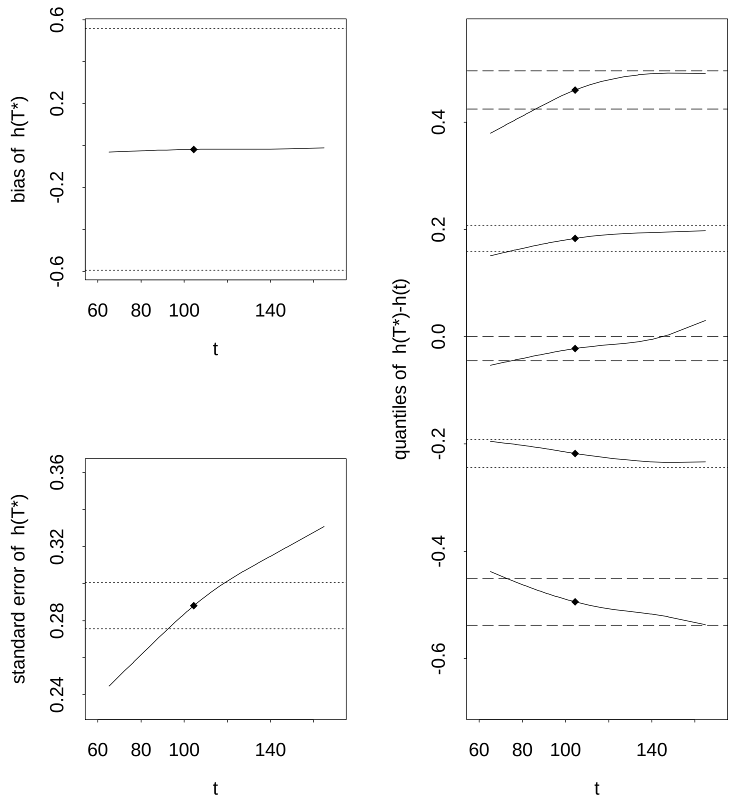

Figure 10: Insurance claim data. As in Figure 9, but for $\log \hat{\xi}_{0.99}$.

5 Resampling Validity

Resampling Models

A critical question in any bootstrap application is whether or not the resampling model is correct. In parametric bootstrap applications this would usually be automatic, but there are exceptions – problems with large numbers of nuisance parameters, for example – as we shall see. In nonparametric bootstrap applications the situation will be less clear-cut.

Qualitatively what we want is for bootstrap samples y^* to behave like real data samples y, which we refer to as the "resample like sample" idea. For example, because the behavior of an estimate T will usually depend upon variability of Y, we would want the variability of Y^* to be comparable to that of Y. Here we look at four non-trivial examples to show how the resample like sample idea can be used.

Example: Count data resampling

Nonhomogeneity is an important factor in Example 7.4 of Davison & Hinkley (1997), which involves modelling of an incomplete two-way table of counts of AIDS cases, in order to predict the numbers

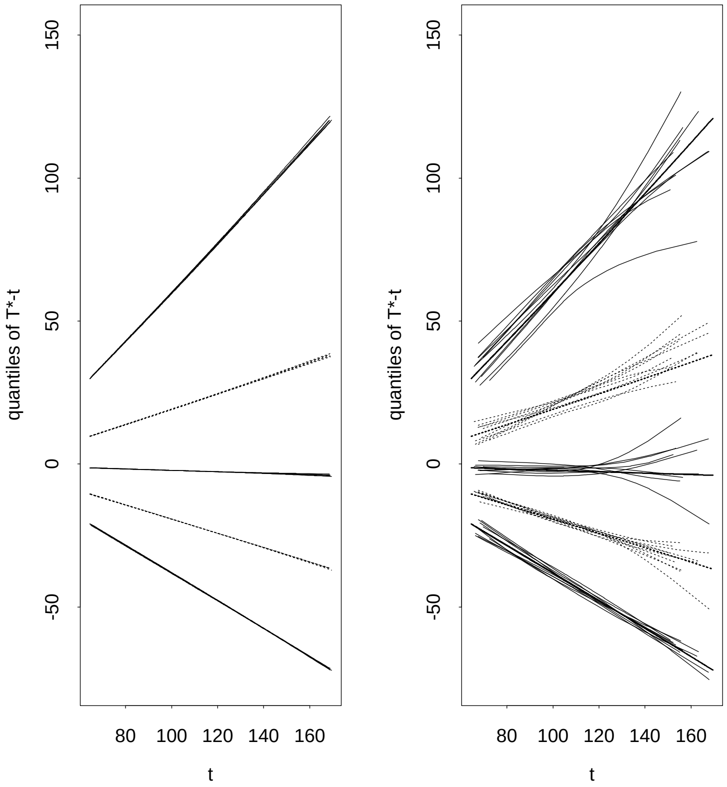

Figure 11: Insurance claim data. As in the right panel of Figure 9, but for ten independent replicates of `pivot.plot` (right panel), and five independent replicates of the double bootstrap analog each with $R = 500$ and $M = 1000$ (left panel).

of diagnoses, subject to reporting delay. There is strong overdispersion relative to Poisson variation, the deviance for a simple log-linear model being 716 on 413 degrees of freedom, and because of the wide range of counts the residuals are very nonhomogeneous, as illustrated in the left panel of Figure 13. This makes simple model-based resampling inappropriate. The right panel of Figure 13 shows boxplots of residual deviances for four resampling schemes, two parametric (Poisson and negative binomial) and two nonparametric (one unstratified and the other stratified) with new datasets generated according to the recipe $y^* = (\hat{\mu} + \hat{\mu}^{1/2}\varepsilon^*)_+$, where ε^* denotes a (mean adjusted) Pearson residual from the original fit and $\hat{\mu}$ a fitted value. The only results that match the original deviance are those for the stratified scheme, which restricts the resampling of residuals within the three strata indicated in the left panel.

In order to assess the usefulness of stratification in this context, we generated 1000 two-way tables of size 20×10 according to a negative binomial model with mean in the (j, k) cell given by

$$\log \mu_{jk} = \beta_0 + \beta_1(j-1)/19 + \beta_2(k-1)/9, \tag{12}$$

where β_0, β_1 and β_2 were chosen so that $\mu_{1,1} = 5$, $\mu_{20,1} = 50$ and $\mu_{1,10} = 0.2$, and with variance

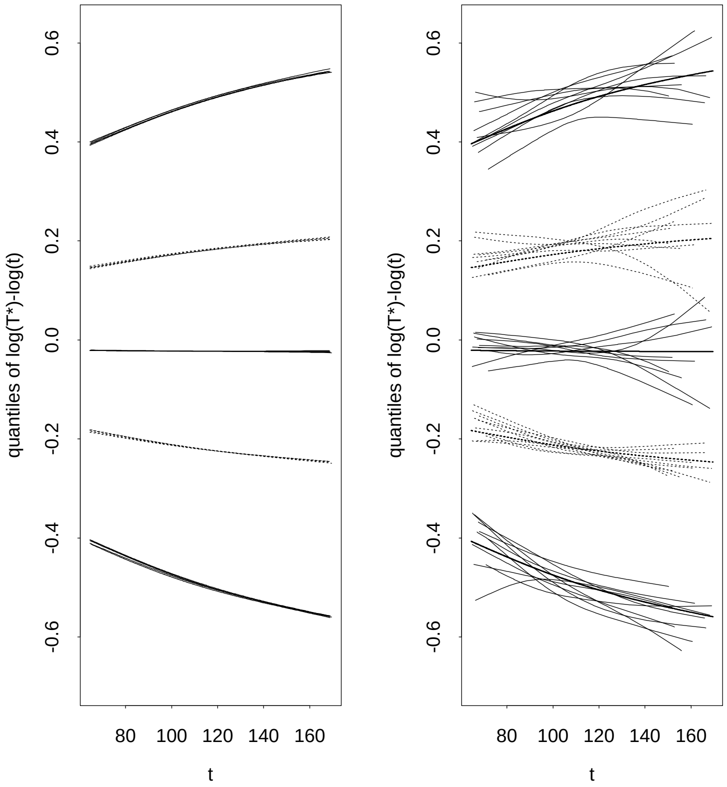

Figure 12: Insurance claim data. As in the right panel of Figure 10, but for ten independent replicates of `pivot.plot` (right panel), and five independent replicates of the double bootstrap analog each with $R = 500$ and $M = 1000$ (left panel).

function $\mu + \mu^2/12$. We then fitted the two-way log-linear model to those cells with $j + k < 22$, i.e. leaving out the lower right corner of the table, and used the fitted model to predict totals $\sum_{k=22-j}^{10} y_{jk}$ for those columns not used in the fitting. Prediction intervals were constructed by treating each

$$Z_j = \sum_{k=22-j}^{10} (y_{jk} - \hat{\mu}_{jk}) / \left(\sum_{k=22-j}^{10} \hat{\mu}_{jk} \right)^{1/2}$$

as pivotal and estimating its quantiles by resampling.

The four resampling schemes used were those described above, namely simulation from fitted Poisson and negative binomial models, nonparametric model-based resampling with no stratification of residuals, and with residuals stratified according to whether $\log \hat{\mu}_{jk}$ lay in the intervals $(-\infty, 1]$, $(1, 2)$ and $[2, \infty)$. For the nonparametric schemes the residuals were mean corrected within strata, so that the means of ε^* are zero, but there was no rounding of the resample values $y^* = (\hat{\mu} + \hat{\mu}^{1/2}\varepsilon^*)_+$, even for the low-count predictions at $j = 12$. Table 3 gives coverages for prediction intervals based on these schemes, for $j = 12$, 16 and 20, corresponding to increasing amounts of extrapolation from the model. The top half of the table corresponds to fitting the model (12) with 9 parameters, while

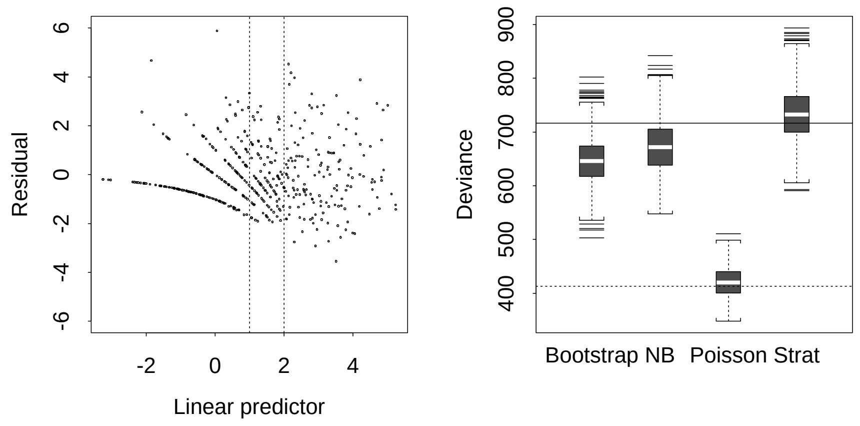

Figure 13: Bootstrap analysis of AIDS data. Left panel: Pearson residuals against linear predictor for log-linear model, with possible strata for resampling indicated by the vertical lines. Right panel: deviances for 999 datasets generated under Poisson, negative binomial, unstratified and stratified nonparametric resampling schemes. The horizontal lines show the original deviance (solid) and the degrees of freedom (dots).

the bottom half corresponds to fitting the additive loglinear model with all row and column effects (29 parameters).

□

The previous example illustrates the relevance of stratification in resampling. The notions of resampling and conditioning may be widely applicable in resampling methods. In this context we think of stratification as a pre-emptive device for making the the total bootstrap output relevant, whereas conditioning is a post-simulation device for selecting only the relevant bootstrap samples, when irrelevant heteroscedasticity or nonhomogeneity is introduced by the resampling scheme. For some discussion of this in the context of regression modelling, see Section 6.3.1 of Davison & Hinkley (1997).

Example: Spatial distribution of caveolae

Consider the data plotted in the left panel of Figure 14, which are spatial locations of caveolae in a small square of muscle tissue. The resampling scheme we have in mind here is tile resampling (Hall, 1985; Davison & Hinkley, 1997, Example 8.9) where the size of tiles must be chosen. The non-Poisson character of the spatial array is measured by a standardized version of Ripley's K-function, defined by

$$K(d) = \lambda^{-1}\mathrm{E}(\text{number of events within distance } d \text{ of an arbitrary event}),$$

Table 3: One-sided coverages (%) of studentized bootstrap prediction intervals for unobserved row totals in 1000 incomplete two-way tables of counts. All bootstrap simulations have $R = 250$.

Nine-parameter model fit

	Row number and target coverage (%)											
	$j = 12$				$j = 16$				$j = 20$			
method	5	10	90	95	5	10	90	95	5	10	90	95
Poisson	0.0	0.0	89.8	94.9	8.3	14.5	85.7	91.9	15.1	22.4	78.4	83.4
Neg Binomial	0.0	0.0	90.5	95.6	4.6	8.3	91.3	96.3	1.9	5.5	93.0	96.9
Nonparametric	0.0	0.7	89.2	95.2	4.9	8.6	88.3	93.3	5.6	10.4	86.8	94.1
Stratified	0.8	12.0	88.0	94.5	9.9	14.4	80.6	87.4	6.8	12.1	85.7	92.0

Twenty-nine parameter model fit

	Row number and target coverage (%)											
	$j = 12$				$j = 16$				$j = 20$			
method	5	10	90	95	5	10	90	95	5	10	90	95
Poisson	0.5	3.2	89.0	95.5	8.3	14.2	83.4	89.1	19.1	25.0	71.2	77.0
Neg Binomial	0.1	1.5	89.2	95.8	5.2	10.2	87.0	93.4	2.5	6.3	87.6	91.3
Nonparametric	0.3	2.0	83.9	88.6	3.9	8.4	83.4	89.2	11.6	18.4	73.2	79.4
Stratified	0.8	6.1	84.7	89.4	4.1	9.1	84.4	91.4	7.7	15.2	80.5	86.4

with λ the marginal density of points. The standardized form is

$$Z(d) = \{K(d)/\pi\}^{1/2} - d,$$

which for a Poisson process is exactly equal to 0 for all radii d. The empirical version $\hat{Z}(d)$ of $Z(d)$ is obtained from an edge-corrected empirical version of $K(d)$, as described in Chapter 8 of Ripley (1981).

If a resampling scheme correctly mirrors the data, then corresponding resample estimates $\hat{Z}^*(d)$ should look like $\hat{Z}(d)$. We examine this graphically by plotting both $\hat{Z}(d)$ and R replicates of $\hat{Z}^*(d)$: the envelope of these replicates might serve just as well. Figure 15 shows the result for four different tile sizes with the caveolae data, and $R = 19$. Only for the 3×3, and possibly 5×5, tiling does the resample envelope capture both the low K-values near $d = 20$ and the near-constant level beyond $d = 50$.

The corresponding plots for a simulated random pattern (binomial process) with the same number of points is shown in Figure 16. Now all levels of tiling appear to be equally good, 15×15 as good as 3×3, in terms of reproducing spatial character. The larger pool of possible resamples provided by 15×15 tiling would tend to give smaller bootstrap simulation error.

□

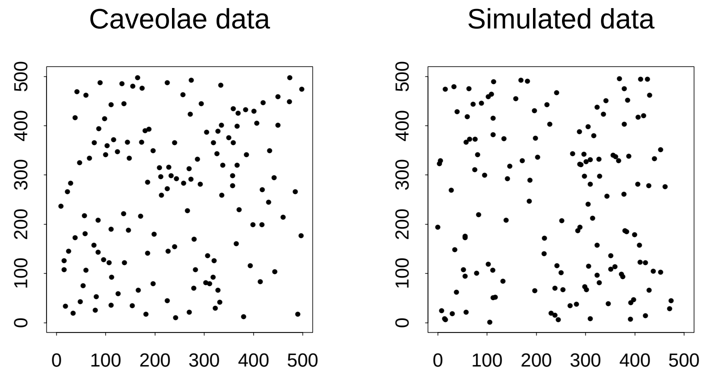

Figure 14: Caveolae data. Left panel shows locations of $n = 137$ points in muscle tissue. Right panel shows the same number of points simulated from a simple binomial process.

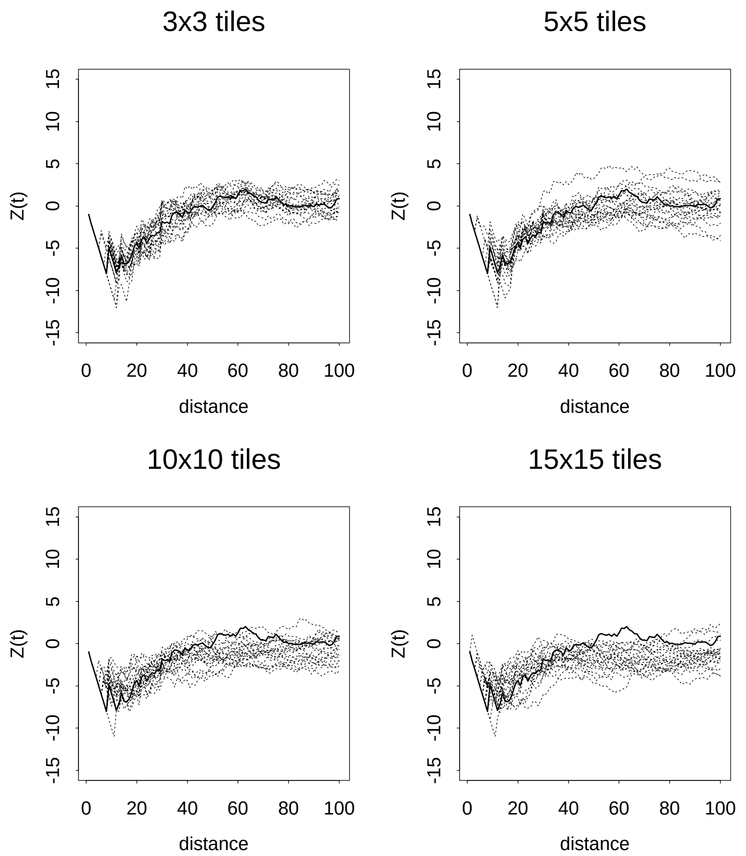

Figure 15: Caveolae data. Standardized Ripley K-functions $Z(t)$ for data (bold curves) and for $R = 19$ resampled patterns (dotted curves) for four levels of tile resampling.

Example: Time series of river heights at Manaus, Brazil

A general strategy for bootstrapping dependent data is block resampling, several variants of which have been proposed and studied for their ability to deal with stationary time series data. A general

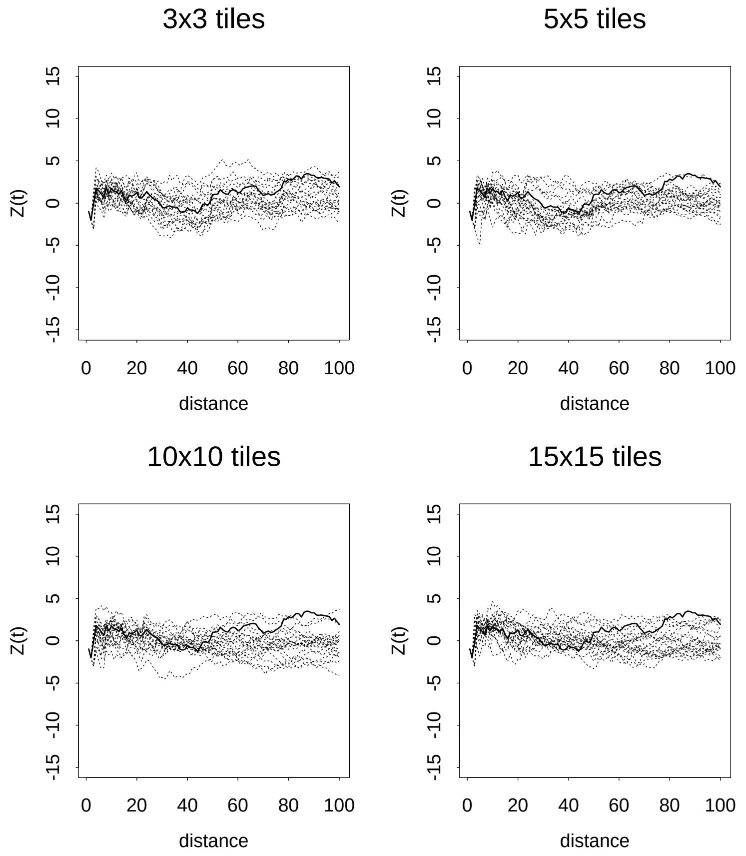

Figure 16: Simulated binomial process data. Standardized Ripley K-functions $Z(t)$ for data (bold curves) and for $R = 19$ resampled patterns (dotted curves) for four levels of tile resampling.

introduction with examples is given in Chapter 8 of Davison & Hinkley (1997), and recent accounts with extensive bibliographies are given by Lahiri (1999) and Lahiri, Kaiser, Cressie and Hsu (1999).

The simplest form of fixed-length block resampling patches together randomly selected blocks of length l: we ignore minor difficulties associated with non-integer n/l. A key question for any particular application is choice of l. To assess whether or not a particular l is suitable, it is useful to plot the sample correlogram, partial correlogram or cumulative periodogram, and then superimpose the corresponding plots for a small number (e.g. $R = 19$) of resampled series.

Figure 17 illustrates this for the Manaus data (Davison & Hinkley, 1997, p. 388) using the cumulative periodogram; the 'best' autoregressive process chosen using AIC is of order $p = 8$. The upper panels correspond to block resampling of the entire series of length 1080 with blocks of length $l = 1$, 3, 5 and 10. For a suitable resampling model we would expect the resample curves to envelope the data curve, but evidently none of these capture the second-order properties of the original data.

One strategy for inducing the appropriate serial dependence in the resampled time series is post-blackening, in which a suitable model is fitted to the series, the corresponding residuals are

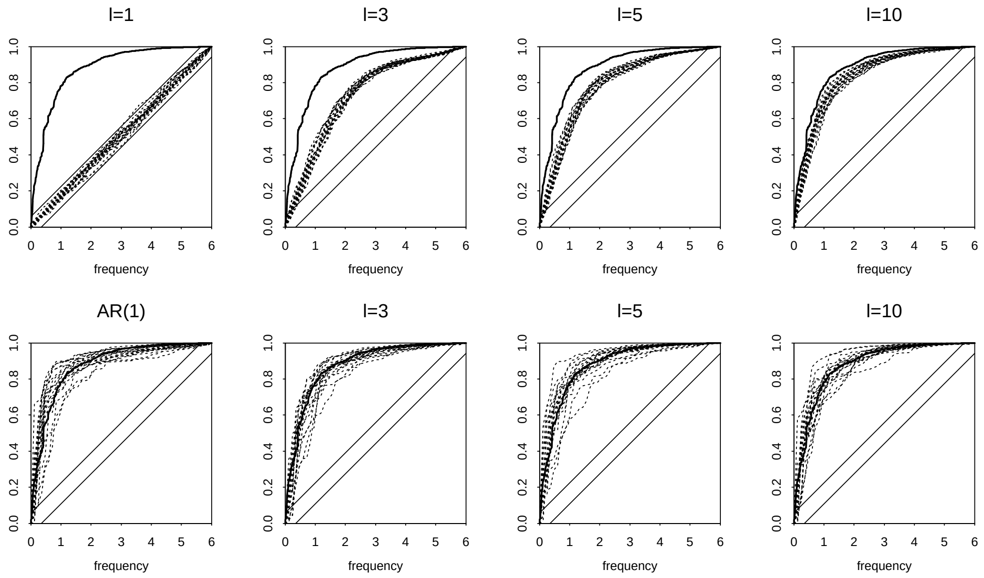

Figure 17: Manaus data. Cumulative periodograms for data (bold curves) and for $R = 19$ resampled series taken in various ways. The upper panels show results for block resampling with block length l, while the lower panels show results for post-blackening using a fitted AR(1) model.

resampled in blocks, and bootstrap time series are generated by applying the estimated model to the resampled residuals. The residuals are generally much closer to white noise than the original observations, so shorter blocks of residuals can be used. This is illustrated in the lower panels of Figure 17, which compare cumulative periodograms for post-blackened series with the original panels above; the model fitted was AR(1). Clearly the resampling results depend less on block length, and it seems that post-blackening with $l = 5$ or 10 should reproduce the second-order structure of the original series.

To judge the efficacy of post-blackening, a series of simulation experiments was performed to compare the coverage of normal confidence intervals for the mean of time series, with bias and variance estimated from $R = 49$ resampled series; the statistic used was the 25% trimmed average. We generated 1000 series of length $n = 100$ from an AR(2) model with parameters 0.8 and -0.3, an MA(2) model with parameters 1 and -0.5 and an ARMA(1,1) model with parameters 0.8 and -0.3; in each case we took t_5 innovations. To each of these series we fitted the best AR(p) model, with $p \leq 5$, as judged by AIC, and performed post-blackening with block lengths $l = 1$, 3, 5 and 10; of course post-blackening with $l = 1$ amounts to resampling innovations independently from residuals. We compared this with ordinary block resampling, using blocks of length $l = 3$, 5 and 10, and with model-based resampling from the best fitted AR(p) model, with $p \leq 5$. All bootstrap simulations used $R = 49$ and the same seed for the random number generator, in order to reduce the standard errors for comparison among procedures.

Table 4 shows the resulting coverages. Post-blackening with $l = 10$ produces excellent coverages for the AR and MA models, and while it is less successful for the ARMA model, it (just!) continues to perform best overall. Post-blackening seems a useful strategy for inducing dependence, and theoretical investigation of its properties would seem worthwhile.

Table 4: One-sided coverages (%) of bootstrap normal confidence intervals for the mean of time series of length 100, based on model-based resampling from a fitted AR(5) model, block bootstraps with block length $l = 3$, 5 and 10, and post-blackening with a fitted AR(p) model, $p \leq 1$ and $l = 1$, 3, 5 and 10. In each case the standard error is less than 1%.

		Target coverage (%)											
		AR(2)				MA(2)				ARMA(1,1)			
method	l	80.0	90.0	95.0	97.5	80.0	90.0	95.0	97.5	80.0	90.0	95.0	97.5
AR(5)		77.0	86.7	91.9	94.0	74.6	82.2	88.9	92.5	74.4	84.0	89.9	92.7
Block	3	78.3	88.0	92.6	95.1	90.5	97.8	99.2	99.6	67.3	75.8	81.8	85.5
	5	79.2	88.9	93.1	95.4	87.4	96.0	98.7	99.3	70.0	79.0	84.9	88.4
	10	78.8	88.3	92.4	95.0	83.6	93.2	96.8	98.4	72.6	82.2	88.6	92.4
Post-	1	83.5	92.3	96.4	98.2	87.6	95.4	98.4	99.4	71.0	81.1	87.1	90.7
black	3	86.2	94.2	97.2	98.7	85.5	94.6	97.6	99.1	70.2	80.4	86.7	90.9
	5	84.5	92.9	96.7	98.1	83.0	92.4	96.4	98.8	71.8	82.0	88.5	92.0
	10	82.2	90.6	94.8	97.0	79.8	90.2	94.8	97.1	73.2	84.3	89.5	92.8

□

Example: Shrinkage estimators in wavelet regression

The substitution principle was seen not to work in the case of the Stein shrinkage estimator in Section 3. This happens largely because of the resampling model. Beran (1995) suggests that a more appropriate parametric resampling model would involve modified mean estimates that are obtained with less severe shrinkage factor than in (3) This modification can be found using the "resample like sample" idea, here interpreted to mean equating $\sum(y_j - \bar{y})^2$ and $\sum(y_j^* - \bar{y}^*)^2$, the latter replaced by its resample model expectation; see Problem 3.18 in Davison & Hinkley (1997). Illustrations of the effect of the modification are shown in the rightmost columns of Figures 5 and 6: the remedy is at best a partial success.

What happens with the Stein estimator is relevant to the general problem of nonparametric curve fitting, where standard methods replace unbiased estimates (perhaps the data values themselves when no replication exists) by smoothed functions of the data. For discussion of kernel smoothing methods see Härdle & Bowman (1988) and Härdle & Marron (1991). Here we look briefly at wavelet regression, where unbiased wavelet coefficient estimates are replaced by super-efficient "thresholded" values: for general introductions see Bruce & Gao (1996) and Silverman (1998). For more detailed discussion than is given here, see Hinkley, Wu & Yang (2000).

To keep the discussion simple, imagine that we observe $y_i = f(x_i) + \varepsilon_i, i = 1, \ldots, n$, for equally-spaced values of x_i and independent homoscedastic errors ε_i, with $n = 2^M$. Then a wavelet estimate for the function $f(x)$ is obtained in three steps:

- **Step 1** Apply the discrete wavelet transform to data $(y_1, \ldots, y_n)$ to obtain unbiased coefficient estimates $c_{j,k}$ for the parameters $\gamma_{j,k}$ and $b_{J,k}$ for parameters $\beta_{J,k}$ in the multiresolution approximation

$$f_{\text{app}}(x) = \sum_{k=1}^{2^{M-J}} \beta_{J,k}\phi_{J,k}(x) + \sum_{j=1}^{J} \sum_{k=1}^{2^{M-j}} \gamma_{j,k}\psi_{j,k}(x)$$

  to $f(x)$, where $\psi_{j,k}(x) = 2^{-j/2}\psi_0\left(2^{-j}x - k\right)$ for a selected "mother" wavelet ψ_0 and $\phi_{J,k}(x) = 2^{-J/2}\phi_0\left(2^{-J}x - k\right)$ for the corresponding "father" wavelet ϕ_0.

- **Step 2** Choose (for example using cross-validation or unbiased estimation of risk) a threshold value h and replace estimates $c_{j,k}$ either by "soft thresholding" values

$$\hat{\gamma}_{j,k}(h) = \text{sign}(c_{j,k})\max(0, |c_{j,k}| - h),$$

  or by "hard thresholding" values

$$\hat{\gamma}_{j,k}(h) = \begin{cases} 0 & \text{if } |c_{j,k}| \leq h, \\ c_{j,k} & \text{if } |c_{j,k}| > h. \end{cases}$$

- **Step 3** Construct estimate $\hat{f}$ by inverse wavelet transform,

$$\hat{f}(x;t) = \sum_{k=1}^{2^{M-J}} b_{J,k}\phi_{J,k}(x) + \sum_{j=1}^{J} \sum_{k=1}^{2^{M-j}} \hat{\gamma}_{j,k}(t)\psi_{j,k}(x).$$

Step 2 is usually applied only at levels $j = J - 2$ and lower, i.e. $\hat{\gamma}_{j,k}(h) = c_{j,k}$ for $j = J, J - 1$ independently of h. (Note that the labelling of levels differs from one author to another in the literature: here we adopt the S-Plus convention that smaller j corresponds to larger "frequency".) "Soft" thresholding appears to be preferred by statisticians. In general h is the realized value of a random variable, but the simple so-called "universal" choice has $h = (2\log n)^{1/2}$. Different threshold values could be chosen at each level j, but for simplicity we shall assume a common value for all levels, namely the preceding universal value. All calculations for this example were done using `S+Wavelets`.

Standard model-based resampling, as described in Chapter 6 of Davison & Hinkley (1997), constructs resamples

$$y_j^* = \hat{f}(x_j;t) + \varepsilon_j^*, \quad j = 1, \ldots, n,$$

with ε_j^* randomly sampled from the n residuals $e_i = y_i - \hat{f}(x_i;t)$. (Leverage correction of residuals would have negligible effect here, because the number of nonzero $\hat{\gamma}_{j,k}(t)$ would be small compared to n.) This resampling procedure is algebraically identical to the "universal" method of resampling in the wavelet transform domain described in Section 6.7 of Bruce & Gao (1996).

Table 5: Averages (times 10^4) of true AMSE and bootstrap $\widehat{\text{AMSE}}$ for wavelet regression with doppler test function, $n = 512$; soft thresholding method. 20 data sets in each case, and $R = 50$ resamples for each data set. snr = signal-to-noise ratio.

	snr = 2	snr = 4
average AMSE	82.7	35.2
average $\widehat{\text{AMSE}}$ with $p = q = 1$	41.4	25.5
average $\widehat{\text{AMSE}}$ with $p = q = 0.25$	86.7	28.1
average $\widehat{\text{AMSE}}$ with $p = q = 0.125$	–	37.6

The resample analog of $\hat{f}$ is obtained by replicating Steps 1–3 above, but using the y^*'s instead of the y's. In general this entails a new threshold value h^* for each resample. Then if, for example, we wish to estimate the average mean squared error of the wavelet curve fit,

$$\text{AMSE} = n^{-1} \sum_{i=1}^{n} \text{E}\{f(x_i) - \hat{f}(x_i; H)\}^2,$$

we take the average over R resample datasets,

$$\widehat{\text{AMSE}} = (Rn)^{-1} \sum_{r=1}^{R} \sum_{i=1}^{n} \{\hat{f}(x_i; h) - \hat{f}^*(x_i; h_r^*)\}^2. \tag{13}$$

Confidence bands for $f(x)$ would most simply connect pointwise upper and lower confidence limits obtained by applying one of the standard bootstrap confidence interval methods, such as BC_a or studentized, at each x_i.

Table 5 summarizes the behavior of $\widehat{\text{AMSE}}$ relative to AMSE for one test application, $n = 512$ observations on the "doppler" curve: this curve and two data sets, with differing levels of noise, are shown in Figure 18. The underestimation by the bootstrap method is appreciable and not unusual: in this case it is due mostly to the low amplitude, high frequency components for small x.

The diagnostic plot, used earlier in Section 3 for the Stein estimator, is applied here to one data set at each noise level, with the results shown in the top panels of Figure 19. The quantities plotted are those defined in (4), here expressed as

$$a^* = \textstyle\sum_{j,k} \left(\hat{\gamma}^*_{j,k} - c^*_{j,k}\right)^2, \quad d^* = \textstyle\sum_{j,k} \left(c^*_{j,k} - c_{j,k}\right)^2.$$

The correlations in the plots suggest the dependence that corresponds to inconsistency of the standard bootstrap.

In the spirit of Beran's suggestion for bootstrapping the Stein estimator, we might consider the modified resampling model

$$y_i^* = \hat{f}(x_i; ph) + \varepsilon_i^*, \quad i = 1, \ldots, n, \tag{14}$$

with ε_i^* randomly sampled from the n residuals

$$e_i = y_i - \hat{f}(x_i; qh), \tag{15}$$

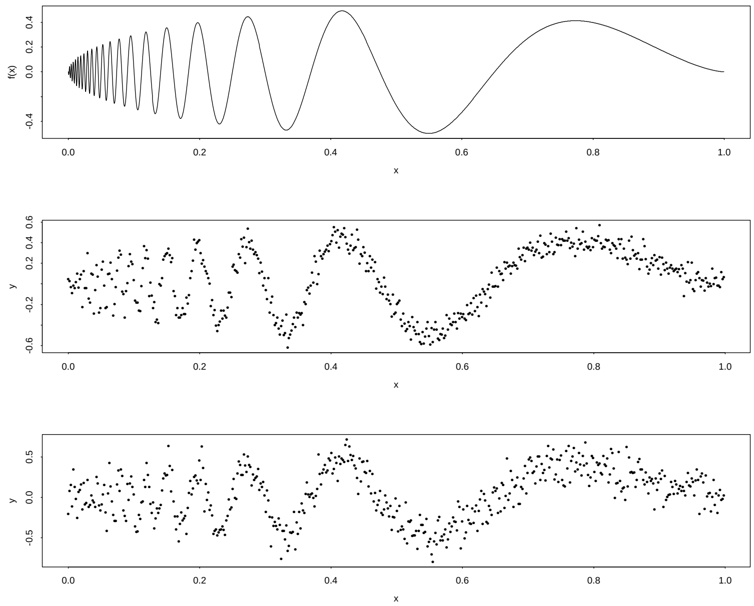

Figure 18: Doppler curve (top panel) and two simulated data sets, $n = 512$, signal-to-noise ratios 4 (middle panel) and 2 (bottom panel).

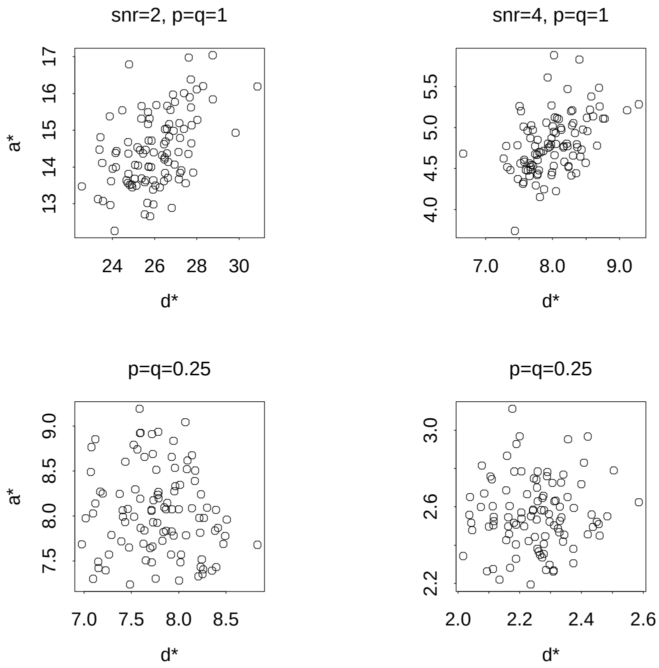

Figure 19: Diagnostic plot for failure of standard bootstrap for wavelet curve fit. Doppler data $n = 512$; signal-to-noise ratios 2 and 4; $R = 100$. Top panels are for naive resampling ($p = q = 1$), bottom panels are for $p = q = 0.25$.

where $0 < p, q \leq 1$. That is, the modified resampling model has systematic part *less* smooth than the estimate $\hat{f}$ and stochastic part *less* variable than the original residuals. In a full discussion of consistent bootstrapping, one would be concerned with the dependence of p and q on n. Here, however, we are concerned only to see whether or not the diagnostic plot could indicate suitability of the modified resampling scheme for particular p and q. To this end, the bottom panels of Figure 19 illustrate the effect for the doppler data examples when $p = q = 0.25$ and $p = q = 0.125$. These plots suggest independence, and hence suitability of the modified bootstrap. As a limited check on this, a small simulation was done to produce the last three rows of Table 5, and this does show reasonable performances of modified resampling schemes for suitably small p and q.

An alternative diagnostic, similar in spirit to that used for the count data example, is an aggregate energy plot for the wavelet fits to data and resamples: the energy plot is a barplot of the squared amplitude of both signal and noise at every level j. Figure 20 illustrates this for the doppler test data with snr $= 2$, and five resamples for each of $p = q = 1$ and $p = q = 0.25$. For practical use one would want to develop one aggregate plot for each resampling model, with a larger value of R.

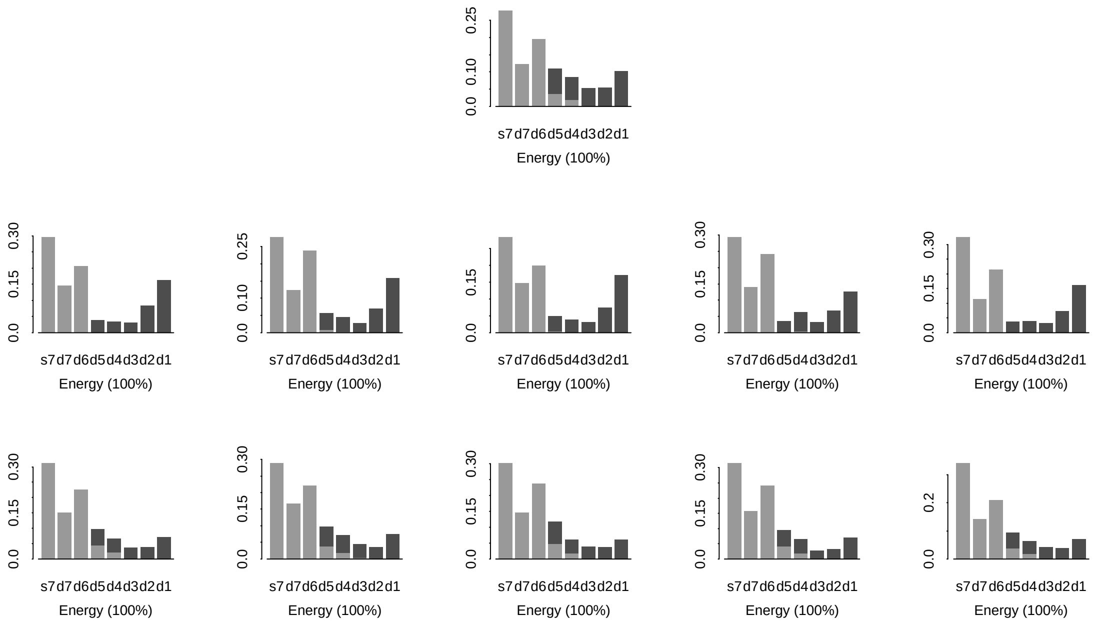

Figure 20: Energy plots for wavelet fit to doppler test data (top row), to five resamples with $p = q = 1$ (middle row), and to five resamples with $p = q = 0.25$ (bottom row). Light shading corresponds to signal, dark shading to noise.

Similar behavior has been observed with other test functions at these signal-to-noise levels; see Hinkley, Wu and Yang (2000).

□

Table 6: Costs (pounds) of cognitive behavioural therapy and the usual therapy for patients with a history of deliberate self harm (Evans, et al., 1999).

cognitive behavioural	30	172	210	212	335	489	651	1263	1294	1875
therapy	2213	2998	4935							
usual therapy	121	172	201	214	228	261	278	279	351	561
	622	694	848	853	1086	1110	1243	2543		

Multiple Sources of Variation

For models with more than one source of variation, such as nested random effect models, it is necessary for data simulation to properly incorporate all sources. In the case of the parametric bootstrap this will usually be straightforward, given data estimates of parameters (although see the earlier discussion of the Stein estimator). The nonparametric bootstrap presents more of a challenge, as is clear from the discussion in Section 3.8 of Davison & Hinkley (1997). There is little more to say about this here, except to note that for any proposed resampling scheme it would be natural to plot histograms of bootstrap residual deviances or residual sums of squares corresponding to each source of variation, checking that these are distributed around the data values of these quantities. In other words, we would produce analogs of plots such as those in Figure 13 for each source of variation.

6 Sensitivity Analysis

One simple but useful approach to assessing the reliability of a bootstrap application is to see how conclusions change when the analysis is performed for several reasonable models. Certain calculations may be less sensitive than others, and in practice it is valuable to know this. We illustrate this with data in Table 6 on the costs of cognitive behavioural therapy and usual care for the treatment of patients with a history of deliberate self harm. The distribution in each group is very skewed, as is typical of cost data, and probability plots suggest that exponential or log normal distributions will fit about equally well. The question of interest is whether the mean costs for the groups are equal. One way to assess this would be to perform a two-sample t test, directly or via confidence intervals, on the difference of the averages $\bar{y}_2 - \bar{y}_1$ for the original data, but the positive and skewed nature of the costs suggests an alternative analysis of the ratio $\bar{y}_2/\bar{y}_1$ or equivalently its logarithm. The studentized bootstrap could then be applied to $t = \log \bar{y}_2 - \log \bar{y}_1$, with estimated variance $v = n_2^{-1} s_2^2/\bar{y}_2^2 + n_1^{-1} s_1^2/\bar{y}_1^2$, where n_k and s_k^2 are the sample sizes and variances. Is this preferable to using the t statistic with the original data, and thereby constructing a confidence interval for the difference of mean costs?

Table 7 shows various 95% confidence intervals obtained with $R = 999$ replicates of parametric bootstraps from fitted exponential and log normal models and from the nonparametric two-sample

Table 7: 95% confidence intervals for difference and ratio of means based on the data in Table 6, using stratified nonparametric resampling, and parametric simulation from fitted exponential and log normal models. All bootstrap simulations have $R = 999$.

		Normal	Percentile	Basic	Studentized
Difference	Nonparametric	$(-1433, 145)$	$(-1499, 102)$	$(-1372, 229)$	$(-1954, 91)$
	Exponential	$(-1385, 107)$	$(-1436, 12)$	$(-1282, 167)$	$(-1758, 82)$
	Log normal	$(-2976, 2747)$	$(-4752, 126)$	$(-1396, 3483)$	$(-2108, 26)$
Ratio	Nonparametric	$(0.232, 1.039)$	$(0.265, 1.159)$	$(0.220, 0.965)$	$(0.217, 1.134)$
	Exponential	$(0.243, 1.030)$	$(0.250, 1.017)$	$(0.251, 1.021)$	$(0.231, 1.177)$
	Log normal	$(0.182, 1.970)$	$(0.113, 1.198)$	$(0.213, 2.256)$	$(0.192, 1.317)$

bootstrap. Results for the nonparametric and exponential models are broadly similar for both difference and ratio, but the log normal results can be very different, particularly for the difference, because the tail of the distribution can give observations much larger than those generated under the other simulation schemes, which badly affect the resampling results. With samples of sizes 13 and 18, it is impossible to distinguish the two parametric models, so the more appropriate tail weight cannot be chosen. An appropriate remedy is to use a confidence interval method that is robust to the model chosen. Studentized confidence intervals are most stable, and when applied to the ratio they seem broadly preferable here.

The double bootstrap can be used to check this conclusion, although it would typically be too burdensome for routine use as a diagnostic. Here we applied it to the studentized difference and ratio, in order to see whether calibration would result in large changes to confidence intervals based on parametric simulation from the fitted log normal and exponential models. We used 1000 outer- and 250 inner-level replicates. For both ratio and difference, changes were small for exponential data. For log normal data the studentized difference needs a large adjustment, but that for the ratio was appreciably smaller, confirming our comments above.

7 Discussion

Of genuine concern must be the fact, seen in many applications, that in quite complex models – to which bootstrap methods are increasingly being applied – *either* the basic bootstrap must be modified so as to produce reliable results, *or* there is no single "basic" bootstrap, but rather a dauntingly large *family* of bootstraps with little theoretical guidance as to which one suits a particular circumstance. In this paper we have taken some initial steps toward making available useful and practicable diagnostics to help identify situations where a problem exists, and in some cases to help indicate what corrective action might be appropriate.

It is generally recognized that diagnostic methods in applied statistics are not unambiguous. This remains true for bootstrap diagnostics, although we have not necessarily teased out the best ways to search for signs of the problems mentioned at the outset. Indeed, in one notable instance – relative inconsistency – a useful empirical diagnostic has proved elusive.

We hope that these brief remarks will act as a challenge to readers to build on our work, particularly to add to our catalog of "problem applications" and corresponding diagnostics.

Appendix 1: Influence Values

Influence value l_j corresponding to case j can be approximated in several ways. The ideal calculation takes the theoretical formula (Davison & Hinkley, 1997, Sections 2.7 and 3.2) for influence $l(y_j, F)$ and substitutes $\hat{F}$ for F: note that F may be a vector of CDFs. The exact jackknife analog is to calculate $(n-1)(t - t_{-j})$ where t_{-j} denotes the estimator T calculated for the data omitting case j. For computational reasons the jackknife influence value is often approximated by $(n-1)(\bar{t}^* - \bar{t}^*_{-j})$: this is used for the rightmost panels in Figure 2. The regression method uses the linear approximation

$$t^* = t(\hat{F}^*) \doteq t(\hat{F}) + n^{-1} \sum_{j=1}^{n} f_j^* l_j \tag{16}$$

applied to all R resamples, so that the l_js are calculated as coefficients in the regression of t^* on the resample frequencies $f_1^*, \ldots, f_n^*$: this is used in the middle panels of Figure 2. It is easy to see that the latter two methods, which are the most generally applicable because they require no theoretical input and use only bootstrap output, are vulnerable to outliers because of the latter's effects on the t^* values.

Appendix 2: Details of the pivot plots

The purpose of a pivot plot is to suggest how the distribution of a quantity such as $T - \theta$ would be affected by changes in the population parameter value θ. Following the bootstrap substitution principle we instead examine how the distribution of $T^* - t$ varies with values t of T. Our pivot plots contain three panels showing aspects of this relationship, namely estimates of the bias and standard error of T^* and of certain quantiles of the distribution of $T^* - t$. Let $\kappa(\theta)$ denote the quantity of interest, be it bias, variance or a specified quantile.

We estimate $\kappa(t)$ for n_t values of t, taken as equally spaced quantiles of the bootstrap output $t_1^*, \ldots, t_R^*$ between the 0.05 and 0.95 quantiles of the t^*s. We have found that $n_t \geq 20$ is needed to approximate the relationship between $\kappa(t)$ and t well. At each t we first construct a distribution on the data $y_1, \ldots, y_n$ for which T^* is centred at t by smoothing the original bootstrap frequencies as in (6) and importance reweighting the y_j using (8) and (9). From this we calcualte the corresponding importance sampling weights to estimate the value of $\kappa(t)$ that would have arisen had we sampled from a distribution centred at t. Of the available methods of importance sampling estimation, we have found that the importance sampling regression estimates (Hesterberg, 1995) perform best for this purpose.

As described above the technique leads to a bias that stems from using the same set of bootstrap frequencies to find both recentred distribution and importance sampling weights. This bias can be reduced by splitting $t_1^*, \ldots, t_R^*$ into two halves, and using one half for the frequency smoothing and the other for the importance sampling and estimation. The roles of the halves are then reversed, giving two estimates of $\kappa(t)$. To ensure that the halves cover similar ranges, the t^*s should be stratified, for example by taking alternate ordered values.

Ventura (1997, 1998) showed that importance sampling estimates in the tails of a bootstrap distribution can suffer from pseudo-bias unless some of the replicates are generated from a distribution centred in the tails. We therefore recommend that $S \geq 100$ additional bootstrap replicates are generated from each of the two distributions centred at the 0.05 and 0.95 quantiles of the original t^*s. We use an exponentially tilted bootstrap to generate these extra replicates, which are used in the importance sampling estimation at each t but not in the frequency smoothing. This also serves to reduce the extreme variability often associated with importance resampling in the tails of a distribution.

Once this process is completed we have $2n_t$ importance sampling estimates of $\kappa(t)$ near n_t values of t. We smooth the $2n_t$ estimates using cubic splines with 3 degrees of freedom. The resulting oversmoothing is useful, because our interest is in any overall trend in $\kappa(t)$, which should be quite smooth in t. To account for the changing variance of the estimates of $\kappa(t)$, we use the reciprocal of the variance of the importance sampling weights as a weight for the cubic spline fits.

If the quantity of interest is truly pivotal, then all the smoothed curves in the plots should be roughly horizontal. To help assess this we add horizontal lines around each curve, showing the limits of an approximate 95% confidence interval at the original value of T. Curves that leave these bounds in a systematic way suggest that the quantity examined is not a pivot.

The first author has implemented this procedure in `S-Plus`, using the library of bootstrap functions released with Davison & Hinkley (1997). The function `pivot.plot` will be included in future releases of the library, but for now is available from A. J. Canty. Specific implementation details are explained in the accompanying help-file.

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada through a grant to the first author, by the Swiss National Science Foundation and by the UK Engineering and Physical Sciences Research Council through various research grants and especially an Advanced Research Fellowship awarded to the second author. Thanks go to Dongfeng Wu, Yuchieh Yang, Julie Barber and Christophe Osinski for help with examples.

References

Beran, R.J. (1995) Stein confidence sets and the bootstrap. *Statistica Sinica*, **5**, 109–127.

Beran, R.J. (1996) Bootstrap variable selection and confidence sets . In *Robust Statistics, Data Analysis and Computer Intensive Methods*, H. Rieder, ed., Springer Lecture Notes in Statistics **109**.

Beran, R.J. (1997) Diagnosing bootstrap success. *Annals of the Institute of Statistical Mathematics*, **49**, 1–24.

Bertail, P., Politis, D.N., and Romano, J.P. (1999) On subsampling estimators with unknown rate of convergence. *J. Amer. Statist. Assoc.*, **94**, 569–579.

Bickel, P.J. and Freedman, D. (1981) Some asymptotic theory for the bootstrap. *Ann. Statist.*, **9**, 1196–1217.

Bruce, A. and Gao, H. (1996) *Applied Wavelet Analysis with S-PLUS.* New York: Springer.

Davison, A.C. and Hinkley, D.V. (1997) *Bootstrap Methods and Their Application.* Cambridge University Press.

Davison, A.C., Hinkley, D.V., and Worton, B.J. (1992) Bootstrap likelihoods. *Biometrika*, **79**, 113–130.

Davison, A.C., Hinkley, D.V., and Worton, B.J. (1995) Efficient and accurate construction of bootstrap likelihoods. *Statistics and Computing*, **5**, 257–264.

Efron, B. (1992) Jackknife-after-bootstrap standard errors and influence functions (with Discussion). *J. Roy. Statist. Soc. B*, **54**, 83–127.

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

Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997) *Modelling Extremal Events for Insurance and Finance.* Berlin: Springer.

Evans, K., Tyrer, P., Catalan, J., Schmidt, U., Davidson, K., Dent, J., Tata, P., Thornton, S., Barber, J. and Thompson, S. (1999) Manual-assisted-cognitive-behaviour therapy in the treatment of recurrent deliberate self harm: A randomised controlled trial. *Psychological Medicine*, **29**, 19–25.

Hall, P. (1985) Resampling a coverage pattern. *Stochastic Processes and Their Applications*, **20**, 231–246.

Hall, P. (1986) On the bootstrap and confidence intervals. *Ann. Statist.*, **14**, 1431–1452.

Härdle, W. and Bowman, A.W. (1988) Bootstrapping in nonparametric regression: local adaptive smoothing and confidence bands. *J. Amer. Statist. Assoc*, **83**, 102–110.

Härdle, W. and Marron, J.S. (1991) Bootstrap simultaneous error bars for nonparametric regression. *Ann. Statist.*, **19**, 778–796.

Hesterberg, T.C. (1995) Weighted average importance sampling and defensive mixture distributions. *Technometrics*, **37**, 185–194.

Hinkley, D.V., Wu, D. and Yang, Y. (2000) Bootstrapping wavelet regression: diagnostics and solutions. University of California at Santa Barbara, Department of Statistics and Applied Probability.

Lahiri, S.N. (1999) Theoretical comparisons of block bootstrap methods. *Annals of Statistics*, **27**, 386–404.

Lahiri, S. N., Kaiser, M. S., Cressie, N. A. C., and Hsu, N.-J. (1999) Prediction of spatial cumulative distribution functions using subsampling (with Discussion). *J. Amer. Statist. Assoc*, **94**, 86–110.

Lee, S.M.S. and Young, G.A. (1995) Asymptotic iterated bootstrap confidence limits. *Annals of Statistics*, **23**, 1301–1330.

Léger, C. and MacGibbon, B. (1998) On the bootstrap and cube root asymptotics. Technical report CRM-2561, Centre de Recherches Mathématiques, Université de Montreal

Politis, D.N. and Romano, J.P. (1994) A general theory for large-sample confidence regions based on subsamples under minimal assumptions. *Annals of Statistics*, **22**, 2031–2050.

Ripley, B.D. (1981) *Spatial Statistics*, New York: Wiley.

Silverman, B.W. (1998) Wavelets in statistics: some recent developments. Unpublished report, University of Bristol.

Singh, K. (1998) Breakdown theory for bootstrap quantiles. *Annals of Statistics*, **26**, 1719–1732.

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

Venables, W.T and Ripley, B.D. (1997) *Modern Applied Statistics in S-Plus, second edition.* New York: Springer.

Ventura, V. (1997). *Likelihood Inference by Monte Carlo Methods and Efficient Nested Bootstrapping.* D.Phil. Thesis, University of Oxford.

Ventura, V. (1998). Nonparametric bootstrap recycling. Technical Report 673, Statistics Department, Carnegie Mellon University.

Wood, A.T.A. (2000) Bootstrap relative errors and sub-exponential distributions. *Bernoulli*, to appear.