Statistical Science
2003, Vol. 18, No. 2, 141–157

Recent Developments in Bootstrap Methodology

A. C. Davison, D. V. Hinkley and G. A. Young

Abstract. Ever since its introduction, the bootstrap has provided both a powerful set of solutions for practical statisticians, and a rich source of theoretical and methodological problems for statistics. In this article, some recent developments in bootstrap methodology are reviewed and discussed. After a brief introduction to the bootstrap, we consider the following topics at varying levels of detail: the use of bootstrapping for highly accurate parametric inference; theoretical properties of nonparametric bootstrapping with unequal probabilities; subsampling and the m out of n bootstrap; bootstrap failures and remedies for superefficient estimators; recent topics in significance testing; bootstrap improvements of unstable classifiers and resampling for dependent data. The treatment is telegraphic rather than exhaustive.

Key words and phrases: Bagging, bootstrap, conditional inference, empirical strength probability, parametric bootstrap, subsampling, superefficient estimator, tilted distribution, time series, weighted bootstrap.

1. INTRODUCTION

Since its introduction by Efron (1979), the bootstrap has become a method of choice for assessing uncertainty in a vast range of domains. So extensive is the literature on the topic that even book-length treatments such as Davison and Hinkley (1997), Shao and Tu (1995), Efron and Tibshirani (1993) or Hall (1992) treat only certain aspects. In this article, we attempt to give a bird's-eye overview of the current state of bootstrap research, treating only sketchily topics that are dealt with elsewhere in this issue and focusing on work that strikes us as most promising for future developments.

A. C. Davison is Professor of Statistics, Institute of Mathematics, Swiss Federal Institute of Technology, 1015 Lausanne, Switzerland (e-mail: Anthony.Davison @epfl.ch). D. V. Hinkley is Professor of Statistics, Department of Statistics and Applied Probability, University of California, Santa Barbara, California 93106–3110 (e-mail: hinkley@pstat.ucsb.edu). G. A. Young is Reader in Methodological Statistics, Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, Cambridge CB3 0WB, United Kingdom (e-mail: G.A.Young@statslab.cam.ac.uk).

Section 2 outlines some basic ideas and in particular describes bootstrap approaches to the fundamental statistical activities of confidence interval construction and testing hypotheses. Subsequent sections review some of the extensions, focusing on theoretical and methodological work that has appeared since publication of the above books. Inevitably, given the enormous volume of published research, our selection of topics is incomplete and of course it reflects our interests. Section 3 describes how bootstrap simulation can be used to provide highly accurate parametric inference and Section 4 focuses on nonuniform nonparametric sampling schemes. Section 5 outlines the topic of subsampling, which generalizes the bootstrap and can repair it when the usual bootstrap is inconsistent. Aspects of bootstrap failure are also discussed in Section 6, which assesses the usefulness of fixing the bootstrap in one important situation where it fails, that is, superefficient estimation at one point of the parameter space. Section 7 gives a brief discussion of developments in bootstrap hypothesis testing and Section 8 outlines how the bootstrap provides a smoothing mechanism that can substantially reduce prediction and classification error in machine learning settings. The final sections touch on dependent data and on further topics.

Although much of our discussion generalizes to more complex situations, for simplicity of exposition we mostly suppose that the data available form a random sample, that is, a set of independent identically distributed random variables.

2. KEY IDEAS

One reason for the success of the bootstrap lies in its simplicity, wherein theoretical understanding can apparently be replaced by repeated computations with random samples. And what a brilliantly chosen name! The most pervasive idea is sometimes called the *substitution principle* or, less loftily, the *plug-in rule*—explicit recognition of the fact that frequentist inference involves the replacement of an unknown probability distribution F by an estimate $\tilde{F}$. In the simplest setting a random sample $Y = (Y_1, \ldots, Y_n)$ is available and the nonparametric estimate is the empirical distribution function $\widehat{F}$, while a parametric model $F(y; \psi)$ with a parameter ψ of fixed dimension is replaced by its maximum likelihood estimate $F(y; \widehat{\psi})$. The choice between parametric and nonparametric estimates depends on the setting, and semiparametric estimates are also in common use, particularly in regression problems. The estimate of F can be modified by the imposition of constraints, as in hypothesis testing problems, or for technical reasons—for instance, to improve a rate of convergence.

Recognizing that the era of cheap computing just around the corner would democratize data analysis, the second idea was to replace analytical calculation of properties of an estimator $\widehat{\theta}$ of an unknown parameter $\theta = \theta(F)$ by simulation from $\tilde{F}$. This gives the familiar generation of R replicate bootstrap samples $\{Y_1^*, \ldots, Y_n^*\}$ by independent sampling from the fitted model $\tilde{F}$ and the use of the corresponding estimates $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_R^*$ to estimate repeated sampling properties of $\widehat{\theta}$. Particularly in nonparametric settings, we refer to such resampling as the ordinary bootstrap. An important point is that, subject to mild conditions, $\theta(\cdot)$ can be the output of an algorithm of almost arbitrary complexity, shattering the naive notion that a parameter is a Greek letter appearing in a probability distribution and showing the possibilities for uncertainty analysis for the complex procedures now in daily use, but at the frontiers of the imagination a quarter of a century ago.

The combination of these two ideas makes the bootstrap a highly flexible tool for inference that is appealing from various viewpoints. It is applicable by nonexpert practitioners in a vast range of applications and yet susceptible to study by theoreticians because of its obvious relationships to existing procedures. Bickel and Freedman (1981) and others investigated conditions under which bootstrap inference is consistent and, in the process, put into place valuable mathematical machinery for studying it. A series of "smoking guns" (e.g., Bretagnolle, 1983) pointing at instances of bootstrap failure spurred researchers to broaden the applicability of the original sampling scheme; we discuss two related approaches to this in Section 5; further recent theoretical discussions are given by Beran (1997) and Putter and van Zwet (1996). An important step forward was the realization (Singh, 1981) that the bootstrap could deliver higher-order accuracy for confidence intervals, equivalent to Edgeworth correction of classical normal intervals, but less painful and less liable to error, and hence more reliable in practice. The subsequent entwining of classical asymptotics and the bootstrap summarized by Hall (1992) owes much to Peter Hall and his coauthors; Hall (1986) was particularly influential.

A good part of the theoretical bootstrap literature concerns the construction of generally reliable nonparametric confidence intervals. The two main approaches to this are based on the construction of Studentized pivots and on the direct use of quantiles of the bootstrap replicates $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_R^*$. The first approach is inspired by the Student t statistic and requires an estimated variance V^* for $\widehat{\theta}^*$ based on the same bootstrap sample. Then an Edgeworth expansion argument shows that the quantiles of $Z^* = (\widehat{\theta}^* - \widehat{\theta})/V^{*1/2}$ provide consistent estimators of the quantiles of $Z = (\widehat{\theta} - \theta)/V^{1/2}$ for a wide class of estimators $\widehat{\theta}$ that fall into the "smooth function model" (Hall, 1992). This results in so-called Studentized bootstrap or bootstrap t confidence intervals for θ that are second-order accurate, that is, the probability that a one-sided interval with nominal level $1-\alpha$ contains θ is $1-\alpha+O(n^{-1})$. This improves on the coverage error for the corresponding normal confidence interval, which is only first-order accurate, differing from the nominal probability by $O(n^{-1/2})$. One view of the second approach is that resampling of $\widehat{\theta}^*$ conditional on $\widehat{\theta}$ is used to approximate sampling from the posterior distribution of θ given $\widehat{\theta}$. This yields two-sided confidence intervals of the form $(\widehat{\theta}^*_{(R\alpha_1)}, \widehat{\theta}^*_{(R(1-\alpha_2))})$, where $\widehat{\theta}^*_{(r)}$ is the rth ordered bootstrap replicate. The simplest and crudest choice, $\alpha_1 = \alpha_2 = \alpha$, yields the percentile intervals, but the corresponding one-sided intervals are only first-order accurate, and improvements have been sought that determine α_1 and α_2 empirically (Efron, 1987; DiCiccio

and Efron, 1992, 1996). The resulting bias-corrected and accelerated (BC_a) intervals and their variants are, like Studentized bootstrap intervals, second-order accurate. Unlike the Studentized intervals, however, BC_a intervals are transformation-invariant. Numerical work has shown that both Studentized bootstrap and BC_a intervals typically show slight undercoverage, although the former intervals perform somewhat better, partly because occasional instability in the variance estimate V can lead to excessively long intervals. Poor coverage of confidence intervals can be improved, sometimes greatly, by a process known as prepivoting (Beran, 1987, 1988), which involves bootstrap correction of bootstrap procedures and usually entails a double or nested bootstrap computation; see Section 4.

The essential elements of a hypothesis test are a null hypothesis H_0 which imposes contraints on the distribution of the data, for example, fixing a mean value, and a test statistic T, large values of which supply evidence against H_0. The degree of disagreement between the data and H_0 is measured by the significance probability or P-value $p_{\text{obs}} = \Pr_0(T \geq t_{\text{obs}})$, where t_{obs} is the value of T actually observed, and the probability is calculated under a null hypothesis distribution. Bootstrap estimation of p_{obs} involves computation under the null hypothesis distribution, usually by simulation from an estimate $\tilde{F}_0$ that satisfies H_0. In the nonparametric case, the null hypothesis may entail changes to the support of $\tilde{F}$, changes to the resampling probabilities attached to $Y_1, \ldots, Y_n$ or some other modification of the empirical distribution function; for related discussion, see the early paragraphs of Section 4. In many comparative test settings the resulting bootstrap tests are almost equivalent to permutation tests, the essential difference being use of sampling with and without replacement.

The need to modify the sampling plan can be eliminated if the test is based on a pivot. Suppose, for example, that H_0 implies that θ equals some fixed value θ_0 and that $(\widehat{\theta} - \theta)/V^{1/2}$ is the basis of the test. Then under the null hypothesis $t_{\text{obs}} = (\widehat{\theta}_{\text{obs}} - \theta_0)/V_{\text{obs}}^{1/2}$ is the observed value of a random variable that has a distribution well approximated by that of $(\widehat{\theta}^* - \widehat{\theta})/V^{*1/2}$ obtained by simulation from either $\tilde{F}_0$ or $\tilde{F}$, because of its pivotality. Thus simulation from a specially constructed null distribution is not needed in this rather special circumstance (Hall and Wilson, 1991). A simpler approach that is sometimes available is to equate inclusion of a null hypothesis value θ_0 in a $1 - \alpha$ confidence interval for θ with $p_{\text{obs}} \geq \alpha$ (Beran, 1986). This coincides with the previous use of a pivot, if that pivot were used for the confidence set. Note, however, that for other than point null hypotheses, care must be taken with the shape of the confidence set; for example, one-sided tests correspond to one-sided intervals with scalar parameters. See also Section 7.

When the data are independent but not identically distributed, the key step in applying the bootstrap is to identify exchangeable components to which resampling can be applied. Often these components are residuals of some sort: if, for instance, a regression model sets $Y_1 = h_1(\psi, \varepsilon_1), \ldots, Y_n = h_n(\psi, \varepsilon_n)$ where the ε_j form a random sample with distribution function G, then G is typically estimated using the empirical distribution function of residuals $\widehat{\varepsilon}_j$ found as the solutions to $Y_j = h_j(\widehat{\psi}, \widehat{\varepsilon}_j)$, where $\widehat{\psi}$ is an estimate of ψ. A bootstrap data set can then be formed by independent sampling of $\varepsilon_1^*, \ldots, \varepsilon_n^*$ from $\widehat{\varepsilon}_1, \ldots, \widehat{\varepsilon}_n$ and setting $Y_j^* = h_j(\widehat{\psi}, \varepsilon_j^*)$. Such model-based bootstrapping may be extended to dependent data situations where a specific parametric model is central to the investigation, for example, in testing whether time series data follow a particular autogressive moving average model. Purely nonparametric bootstrapping is difficult in such circumstances, and parametric or semiparametric models are generally needed. In some cases, stratified resampling is needed, where the strata are empirically chosen to have roughly homogeneous variation. Perhaps the most extreme form of this is the "wild" bootstrap (Wu, 1986; Härdle, 1989, 1990; Mammen, 1993), in which each residual defines its own stratum.

Davison and Hinkley (1997) gave extensive coverage of the topics sketched above, with references to the primary literature.

3. BOOTSTRAPS FOR PARAMETRIC LIKELIHOOD INFERENCE

While the predominant focus in the bootstrap literature has been the development of procedures for accurate nonparametric inference, it has been recognized for some time (DiCiccio and Romano, 1995) that commonly used bootstrap procedures such as the BC_a confidence limit method offer second-order accuracy when applied in parametric models. For parametric inference, the "bootstrap era" has also been a "likelihood era," leading from Efron and Hinkley (1978), Barndorff-Nielsen and Cox (1979) and several articles in the 1980 volume of *Biometrika* (Barndorff-Nielsen, 1980; Cox, 1980; Durbin, 1980; Hinkley, 1980) through Barndorff-Nielsen (1983) to a large literature on second-order accurate likelihood-based inference, much of it synthesized in Barndorff-Nielsen

and Cox (1994) and Severini (2000). More recently, parametric bootstrap schemes have been developed which capture this accuracy automatically.

Let $Y = (Y_1, \ldots, Y_n)$ be a continuous random vector that has a probability density function $f_Y(y; \theta)$ that belongs to some specified parametric family, depending on an unknown vector parameter $\theta = (\gamma, \xi)$ partitioned into a scalar parameter γ of interest and a nuisance parameter ξ. In this setting, inference about γ is typically based on the profile log-likelihood $\ell_{\rm p}(\gamma) = \ell(\gamma, \widehat{\xi}_\gamma)$ and the associated likelihood ratio statistic $w_{\rm p}(\gamma) = 2\{\ell_{\rm p}(\widehat{\gamma}) - \ell_{\rm p}(\gamma)\}$; here $\ell(\gamma, \xi) = \log f_Y(y; \gamma, \xi)$ is the log-likelihood function, $\widehat{\theta} = (\widehat{\gamma}, \widehat{\xi})$ is the overall maximum likelihood estimator of θ and $\widehat{\xi}_\gamma$ is the constrained maximum likelihood estimator of ξ for given γ. Whereas the interest parameter is scalar, inference may be based on the signed root likelihood ratio statistic $r_{\rm p}(\gamma) = \mathrm{sgn}(\widehat{\gamma} - \gamma) w_{\rm p}(\gamma)^{1/2}$. In regular cases, $w_{\rm p}$ has the chi-squared distribution χ^2_1 to error of order $O(n^{-1})$ and $r_{\rm p}$ has the standard normal distribution $N(0, 1)$ to error of order $O(n^{-1/2})$.

A substantial literature that originated with Barndorff-Nielsen (1986) concerns analytically adjusted versions of $r_{\rm p}$ that have the form

$$r_{\rm a} = r_{\rm p} + r_{\rm p}^{-1} \log(u_{\rm p}/r_{\rm p})$$

and that are distributed as $N(0, 1)$ to error of order $O(n^{-3/2})$. The statistic $u_{\rm p}$ depends on specification of an ancillary statistic, a function of the minimal sufficient statistic which is approximately pivotal. The accuracy of χ^2_1 and $N(0, 1)$ approximations to the distributions of $w_{\rm p}$ and $r_{\rm p}$ can be enhanced by parametric bootstrapping. Let Y^* denote a random vector whose density is the fitted parametric density $f_Y(y; \widehat{\gamma}, \widehat{\xi})$ and let $w^*_{\rm p}$ and $r^*_{\rm p}$ denote the versions of $w_{\rm p}$ and $r_{\rm p}$ based on Y^*. It was shown by Martin (1990) that the distribution of $w^*_{\rm p}$ approximates the true distribution of $w_{\rm p}$ to error of order $O(n^{-3/2})$, an improvement over the $O(n^{-1})$ accuracy offered by the χ^2_1 approximation, while Bickel and Ghosh (1990) found that this bootstrap approximation automatically yields Bartlett correction of the likelihood ratio statistic. DiCiccio and Romano (1995) established that the distribution of $r^*_{\rm p}$ under this parametric bootstrap scheme approximates the true distribution of $r_{\rm p}$ to error of order $O(n^{-1})$. Simple parametric bootstrapping therefore provides higher-order accuracy than asymptotic approximation.

EXAMPLE 1 (Exponential regression). Consider an exponential regression model in which lifetimes $T_1, \ldots, T_n$ are independent and exponentially distributed, with means of the form $E(T_i) = \exp(\beta + \xi z_i)$, where $z_1, \ldots, z_n$ are known covariates. Suppose that inference is required for the mean lifetime for covariate value z_0, that is, $\exp(\beta + \xi z_0)$, and let the interest parameter be $\gamma = \beta + \xi z_0$, with nuisance parameter ξ. The signed root likelihood ratio statistic is

$$r_{\rm p}(\gamma) = \mathrm{sgn}(\widehat{\gamma} - \gamma) \cdot \left[2n \left\{ (\gamma - \widehat{\gamma}) + (\widehat{\xi}_\gamma - \widehat{\xi})\bar{c} + n^{-1} e^{-\gamma} \sum_{i=1}^{n} T_i e^{-\widehat{\xi}_\gamma c_i} - 1 \right\} \right]^{1/2},$$

where $\bar{c}$ is the average of $c_i = z_i - z_0, i = 1, \ldots, n$. The analytic calculations leading to the adjusted version $r_{\rm a}$ of $r_{\rm p}$ (Barndorff-Nielsen, 1986) are here readily performed. A one-sided confidence set $(\widehat{\gamma}_1, \infty)$ of coverage $1 - \alpha + O(n^{-1/2})$ is $\{\gamma : r_{\rm p}(\gamma) \le z_{1-\alpha}\}$, with the lower confidence limit $\widehat{\gamma}_1$ obtained by solving $r_{\rm p}(\gamma) = z_{1-\alpha}$, where $z_{1-\alpha} = \Phi^{-1}(1 - \alpha)$ and Φ is the standard normal distribution function. The $O(n^{-1/2})$ coverage error of this confidence set is reduced to $O(n^{-3/2})$ for the corresponding set $\{\gamma : r_{\rm a}(\gamma) \le z_{1-\alpha}\}$ based on the adjusted quantity $r_{\rm a}$. A bootstrap confidence set of nominal coverage $1 - \alpha$ is $\{\gamma : r_{\rm p}(\gamma) \le \widehat{c}_{1-\alpha}\}$ where $\widehat{c}_{1-\alpha}$ denotes a bootstrap estimate of the $1 - \alpha$ quantile of $r_{\rm p}$, obtained as the $1 - \alpha$ quantile of $r_{\rm p}(\widehat{\gamma})$, under sampling from the model with parameter value $(\widehat{\gamma}, \widehat{\xi})$.

In this example, $r_{\rm p}$ is exactly pivotal, so the bootstrap yields the true sampling distribution. Thus the bootstrap confidence set has coverage exactly $1 - \alpha$, rather than a coverage error of order $O(n^{-1})$, as is the case generally. In practice, this set must be constructed by Monte Carlo simulation; a few thousand bootstrap samples are typically needed to make simulation variability negligible.

For numerical illustration we consider data extracted from Example 6.3.2 of Lawless (1982). The $n = 5$ responses T_j are 156, 108, 143, 56 and 1, survival times in weeks of patients suffering from leukaemia, and the corresponding covariates are 2.88, 4.02, 3.85, 3.97 and 5.0, the base-10 logarithms of initial white blood cell count. We take the parameter of interest to be the mean lifetime for $z_0 = \log_{10}(50{,}000)$. For these data, $\widehat{\gamma} = 2.399$ and $\widehat{\xi} = -2.364$. We consider the coverage properties of the three confidence sets for samples of size five from an exponential regression model with these parameter values and the fixed covariate

TABLE 1
Coverages (%) of confidence sets $(\widehat{\gamma}_1, \infty)$ *for mean* $\gamma = \exp(\beta + \xi z_0)$ *at* $z_0 = \log_{10}(50{,}000)$ *in Example* 1, *estimated from* 20,000 *data sets of size* $n = 5$ *and using* $R = 1999$ *bootstrap replicates*

Nominal	1.0	2.5	5.0	10.0	90.0	95.0	97.5	99.0
$\Phi(r_{\mathrm{p}})$	1.5	3.6	6.7	12.8	93.3	96.8	98.6	99.4
$\Phi(r_{\mathrm{a}})$	1.0	2.6	5.4	10.4	89.8	94.9	97.4	99.0
Bootstrap	1.0	2.5	5.1	10.0	89.9	94.8	97.4	98.9

values above. Table 1 compares actual and nominal coverages provided by the three constructions based on 20,000 simulated data sets. Coverages based on normal approximation to r_{p} are quite inaccurate, but normal approximation to r_{a} provides much more accurate confidence sets. Bootstrap confidence sets, here based on 1999 bootstrap samples, have coverages close to nominal levels.

Unless the signed root statistic is exactly pivotal, the bootstrap procedure does not provide exact inference. An important recent development is the observation that the accuracy of inferences based on r_{p} can be further increased by a simple modification of the parametric bootstrap scheme, involving simulation from the model corresponding to the constrained estimator $(\gamma, \widehat{\xi}_\gamma)$ rather than $(\widehat{\gamma}, \widehat{\xi})$. Let $r_{\mathrm{p}}^\dagger$ be the version of r_{p} based on a random vector $Y^\dagger$ that has density $f_Y(y; \gamma, \widehat{\xi}_\gamma)$. DiCiccio, Martin and Stern (2001) showed that approximation of the distribution of r_{p} by that of $r_{\mathrm{p}}^\dagger$ is accurate to order $O(n^{-3/2})$. Under this approach, a confidence set of nominal coverage $1-\alpha$ for the parameter γ of interest is $\{\gamma : r_{\mathrm{p}}(\gamma) \leq c_{1-\alpha}(\gamma, \widehat{\xi}_\gamma)\}$, where $c_{1-\alpha}(\gamma, \widehat{\xi}_\gamma)$ denotes the $1-\alpha$ quantile of the sampling distribution of $r_{\mathrm{p}}^\dagger$, the $1-\alpha$ quantile of the distribution of $r_{\mathrm{p}}(\gamma)$ when the true parameter value is $(\gamma, \widehat{\xi}_\gamma)$. It might appear that such a modified bootstrap confidence set requires a more elaborate Monte Carlo construction. In searching for the endpoint of the set, a separate bootstrap simulation is apparently required for each candidate value of γ. However, this can be avoided by use of the Robbins–Monro stochastic search algorithm (Garthwaite and Buckland, 1992; Carpenter, 1999), under which a single bootstrap sample is generated at each value of γ. Full details of the search procedure in the context of inverting tests to provide nonparametric confidence intervals were given by Garthwaite and Buckland (1992); see also Lee and Young (2003). In simple situations, less sophisticated search procedures are feasible.

The modified parametric bootstrap approach offers the same level of accuracy as provided by adjustments to r_{p} such as r_{a}, while avoiding analytical calculation and specification of ancillary statistics, which is difficult outside restricted parametric classes. The reduction of the order of error, from $O(n^{-1/2})$ to $O(n^{-3/2})$, is a consequence of special properties of the signed root statistic r_{p} which are not enjoyed by inference quantities such as pivots based on Studentization of $\widehat{\gamma} - \gamma$ or the profile score function. In parametric contexts there is therefore a strong theoretical argument in favor of bootstrapping r_{p}.

EXAMPLE 2 (Normal distributions with common mean). Consider inference for the mean, based on a series of independent normal samples with the same mean but different variances. In this example the adjusted quantity r_{a} is intractable, although various messy analytical approximations to it are available.

Example 7.15 of Severini (2000) considered measurements of the strength of six samples of cotton yarn. We model these as $Y_{ij} \overset{\text{ind}}{\sim} N(\gamma, \sigma_i^2)$, $i = 1, \ldots, 6$, $j = 1, \ldots, 4$, and take the common mean γ as the parameter of interest, with orthogonal nuisance parameter $\xi = (\sigma_1, \ldots, \sigma_6)$. We consider the coverages of various confidence sets for 20,000 data sets generated from this model with the parameter values fixed as the maximum likelihood estimates (MLEs) $(\widehat{\gamma}, \widehat{\xi})$ for the data. Table 2 shows the coverages for confidence sets based on normal approximation to r_{p}, normal approximation to an approximation $\tilde{r}_{\mathrm{a}}$ to r_{a} based on orthogonal parameters (Severini, 2000), the simple bootstrap scheme which estimates the sampling distribution of r_{p} by its distribution at parameter value $(\widehat{\gamma}, \widehat{\xi})$ and switching the point of bootstrapping to the constrained maximum likelihood estimator $(\gamma, \widehat{\xi}_\gamma)$. Here the computational burden of construction of the latter intervals is slight, and a simple search procedure, involving drawing 1999 bootstrap samples at each of a set of values of γ, is quite feasible.

Table 2 confirms that the simple bootstrap approach is an improvement over asymptotic inference based on r_{p}. Further substantial gains are obtained by the constrained bootstrap, which improves noticeably on analytical approximation to the r_{a} adjustment.

The analytic approach based on r_{a} is typically highly accurate when the dimensionality of the nuisance parameter is small and r_{a} is easily constructed. The argument for using the modified bootstrap approach then rests primarily on maintaining accuracy while avoiding analytic calculation. In more complex situations, and

TABLE 2
Coverages (%) of confidence intervals for common mean in Example 2, estimated from 20,000 data sets with bootstrap size $R = 1999$

Nominal	1.0	2.5	5.0	10.0	90.0	95.0	97.5	99.0
$\Phi(r_{\rm p})$	3.7	6.7	10.6	16.5	83.2	89.0	93.0	96.1
$\Phi(\tilde{r}_{\rm a})$	1.7	3.7	6.7	12.3	87.3	93.0	96.1	98.1
MLE bootstrap	0.8	2.1	4.4	9.3	90.5	95.5	97.8	99.1
Constrained MLE bootstrap	0.9	2.3	4.7	9.8	89.9	95.1	97.4	99.0

especially when the nuisance parameter is high dimensional and $r_{\rm a}$ must be approximated, the modified bootstrap approach is typically preferable, both in terms of ease of implementation and accuracy.

In addition to the goal of refined distributional approximation, recent developments in parametric likelihood theory are motivated by the Fisherian proposition that inference should be conditional on an appropriate ancillary statistic, when this exists; see, for example, Chapter 2 of Barndorff-Nielsen and Cox (1994) or Chapter 12 of Davison (2003). This means restricting resampling to those samples drawn from a fitted model with ancillary statistic values close to the original sample value. This can be done in Example 1, but numerical results support the view expressed by DiCiccio, Martin and Stern (2001) that when $r_{\rm a}$ is easily constructed, it is likely to be preferable to bootstrapping in terms of conditional accuracy. However, a full evaluation of parametric bootstrap methods in terms of such considerations remains to be undertaken. One approach to conditional parametric bootstrapping in such cases is through the Metropolis–Hastings algorithm (Brazzale, 2000).

4. WEIGHTED NONPARAMETRIC BOOTSTRAPPING

The ordinary nonparametric bootstrap uses uniform resampling from a data sample to mimic the mechanism that originally produced that sample. As noted in Section 2, an exception to this occurs with significance tests, because critical values or significance levels for a test statistic T are determined by its null distribution. For a null hypothesis $H_0 : \theta = \theta_0$, say, the corresponding null distribution F_0 of Y must satisfy $\theta(F_0) = \theta_0$, and in the nonparametric case an appropriate estimate of F_0 is the nonuniform distribution $\tilde{F}_0$ with the same support as $\tilde{F}$ but with the constraint that $\theta(\tilde{F}_0) = \theta_0$. Resampling from $\tilde{F}_0$ is referred to as weighted nonparametric bootstrapping.

Closer attention has recently been paid to resampling with nonuniform probabilities. These weighted nonparametric bootstrap schemes involve a range of bootstrap distributions that depend on a parameter of interest, rather than a single bootstrap distribution, and have much in common with the parametric procedures discussed in Section 3. They encompass a variety of statistical procedures and are related to empirical and other forms of nonparametric likelihood (Owen, 1988; DiCiccio and Romano, 1990). In addition to hypothesis testing, applications include confidence set construction, variance stabilization, nonparametric curve estimation, nonparametric sensitivity analysis and robustification of nonparametric inference through outlier trimming; see Hall and Presnell (1999a, b, c). A general theory which elucidates the effectiveness of weighted bootstrapping in error reduction was given by Lee and Young (2003).

Suppose that $\theta = \theta(F)$ is one parameter of interest and that we wish to estimate F nonparametrically under the constraint $\theta(F) = \theta_0$, where θ_0 may not be the true value of θ. For an arbitrary probability vector $p = (p_1, \ldots, p_n)$, let $\widehat{F}_p$ be the distribution which attaches probability weight p_i to Y_i. Given θ_0 and a data set Y, we choose $p \equiv p(\theta_0)$ to minimize the Kullback–Leibler distance between $\widehat{F}_p$ and $\widehat{F}$,

$$\int \log\{d\widehat{F}(x)/d\widehat{F}_p(x)\}\, d\tilde{F}(x) = -n^{-1} \sum_{j=1}^{n} \log(np_j),$$

subject to the constraint $\theta(\widehat{F}_p) = \theta_0$. The weighted bootstrap uses the resulting $\widehat{F}_p$ as the resampling distribution. We denote samples from this by $Y^\dagger$ to distinguish them from Y^* which is generated from the uniform resampling distribution $p_j \equiv n^{-1}$. Any other parameter $\psi = \psi(F)$ has weighted nonparametric estimator $\widehat{\psi}_0$ under the constraint $\theta = \theta_0$.

Hall and Presnell (1999a) showed theoretical advantages of weighted bootstrapping in specific examples. Here we sketch a general theory due to Lee and

Young (2003) which details the accuracy advantages of weighted over uniform bootstrapping. Consider the effectiveness of bootstrapping in transforming a function $U = u(Y, \theta)$ of the data sample Y and the unknown parameter θ into an approximate pivot, more specifically, an approximately uniform random variable on $(0, 1)$. The error properties of different bootstrap schemes can be assessed by measuring closeness to uniformity, the fundamental goal of bootstrapping being viewed as a transformation of U to a function $U_1 = u_1(Y, \theta)$ say, which is exactly uniformly distributed and so provides exact inference. This process, termed *prepivoting* by Beran (1987), can be applied in particular with U a confidence set root or a test statistic.

Let a one-sided confidence set for θ of nominal coverage $1 - \alpha$ be $\{\theta : u(Y, \theta) \le 1 - \alpha\}$. An example is the percentile method, for which $u(Y, \theta) = \Pr^*(\widehat{\theta}^* > \theta)$, where the asterisk indicates uniform bootstrapping from Y. Other initial roots included in this formulation are those based on normal approximation to the distribution of $\widehat{\theta}$, in which case a confidence set of asymptotic coverage $1 - \alpha$ can be defined by $u(Y, \theta) = \Phi\{(\widehat{\theta} - \theta)/V^{1/2}\}$.

If the sampling distribution of $u(Y, \theta)$ were exactly $U(0, 1)$, then the confidence set would have coverage exactly equal to the nominal coverage $1 - \alpha$. If the distribution is not uniform, there is a discrepancy between the nominal and actual coverages under repeated sampling. By bootstrapping, we hope to produce a new root U_1 so that the associated confidence set $\{\theta : u_1(Y, \theta) \le 1 - \alpha\}$ has lower coverage error.

The uniform bootstrap estimates the distribution function $G(x|\theta)$ of $u(Y, \theta)$ by

$$\widehat{G}(x) = \Pr^*\{u(Y^*, \widehat{\theta}) \le x\},$$

from which we can define the prepivoted root $\widehat{u}_1(Y, \theta) = \widehat{G}\{u(Y, \theta)\}$ for each possible value θ. However, this assumes $u(Y, \theta)$ is exactly pivotal, albeit not exactly $U(0, 1)$. By contrast, the weighted bootstrap with samples $Y^\dagger$ generated from the distribution $\widehat{F}_p$ closest to $\tilde{F}$ in terms of Kullback–Leibler distance, subject to $\theta(\widehat{F}_p) = \theta$, gives the estimate

$$\tilde{G}(x|\theta) = \Pr^\dagger\{u(Y^\dagger, \theta) \le x\}, \tag{1}$$

leading to the weighted prepivoted root $\tilde{u}_1(Y, \theta) = \tilde{G}\{u(Y, \theta)|\theta\}$.

Lee and Young (2003) showed that if $u(Y, \theta)$ is uniform to order $O(n^{-j/2})$, that is,

$$\Pr\{u(Y, \theta) \le u\} = u + O(n^{-j/2}),$$

then under mild conditions the uniform bootstrap root $\widehat{u}_1(Y, \theta)$ is uniform to order $O(n^{-(j+1)/2})$, while the weighted bootstrap root $\tilde{u}_1(Y, \theta)$ is uniform to order $O(n^{-(j+2)/2})$. Thus uniform bootstrapping reduces error by an order of $O(n^{-1/2})$, but weighted bootstrapping is more effective, the error being reduced by $O(n^{-1})$.

EXAMPLE 3 (Folded normal mean). We undertook a simulation study to compare the coverage properties of bootstrap confidence sets for the mean μ when F is folded standard normal, for which $\mu = 0.798$. For 20,000 data sets with $n = 20$, we compared the coverage properties of one-sided confidence sets of nominal coverage $1 - \alpha$ based on the root $u(Y, \mu) = \Phi\{(\widehat{\mu} - \mu)/V^{1/2}\}$, where $\widehat{\mu}$ is the sample mean and $V = \widehat{\sigma}^2/n$, with $\widehat{\sigma}^2$ being the sample variance, and its unweighted and weighted prepivoted forms $\widehat{u}_1(Y, \mu)$ and $\tilde{u}_1(Y, \mu)$, respectively. Intervals constructed from $\widehat{U}_1$ were based on 1999 bootstrap samples. A simple search algorithm was used to construct confidence sets derived from $\tilde{U}_1$, with 1999 weighted bootstrap samples being drawn for each of a set of values of μ, to solve the equation $\tilde{u}_1(Y, \mu) = 1 - \alpha$ which identifies the confidence set limit. Computational efficiency can be improved through the Robbins–Monro procedure; see Section 3. Table 3 displays the coverages of the three intervals: those based on $u(Y, \mu)$ are quite inaccurate, and substantial improvement is given by both conventional and weighted bootstrapping. However, whether the weighted bootstrapping does, as asymptotic theory suggests, outperform conventional bootstrapping depends on the required nominal level $1 - \alpha$.

Lee and Young (2003) showed that this general conclusion applies to regression settings and robust inference, as well as to more conventional problems within the smooth function model. They showed also how iteration of weighted bootstrap prepivoting accelerates the rate of convergence of the error of the bootstrap inference relative to conventional bootstrapping. Such

TABLE 3

Coverages (%) of bootstrap confidence sets for the mean μ of the folded standard normal distribution, estimated from 20,000 data sets of size $n = 20$ and using $R = 1999$ bootstrap replicates; the root taken is $u(Y, \mu) = \Phi\{(\widehat{\mu} - \mu)/V^{1/2}\}$ with $\widehat{\mu} = \bar{Y}$

Nominal	1.0	2.5	5.0	10.0	90.0	95.0	97.5	99.0
$u(Y, \mu)$	4.7	7.6	11.3	17.5	94.5	97.8	99.1	99.7
$\widehat{u}_1(Y, \mu)$	1.8	3.5	6.6	12.9	93.8	97.4	99.0	99.7
$\tilde{u}_1(Y, \mu)$	3.2	5.3	8.6	14.6	93.3	97.0	98.7	99.5

conclusions hold for a whole class of distance measures which generalize the Kullback–Leibler distance (Corcoran, 1998; Baggerly, 1998), allowing the construction of weighted bootstrap distributions $\widehat{F}_p$ that use well-developed algorithms (Owen, 2001) to reduce the added computational burden of weighted bootstrapping.

The same theory applies to testing. For example, when testing a null hypothesis $H_0: \theta = \theta_0$, a one-sided test of nominal size α rejects the hypothesis if $u(Y, \theta_0) \leq \alpha$. If $u(Y, \theta_0)$ were exactly $U(0, 1)$ when $\theta = \theta_0$, then the null rejection probability would be exactly α. To increase the accuracy of an initial root, (1) applied with $\theta = \theta_0$ reduces error by $O(n^{-1})$. In this case, weighted bootstrapping need only be done with the single value θ_0, so computation is no more expensive than uniform bootstrapping. The situation is more complicated for a set null hypothesis.

5. SUBSAMPLING AND THE *m* OUT OF *n* BOOTSTRAP

Bootstrap procedures possess compelling second-order accuracy properties in many settings, but in others they are inconsistent unless problem-specific regularity conditions hold. Thus the development of a subsampling methodology that provides asymptotic consistency under extremely weak conditions, especially where the conventional bootstrap fails, is an important contribution, of which Politis, Romano and Wolf (1999) gave a full account. Its basis is the calculation of a statistic for subsamples of the available data, selected without replacement, and use of these subsample values to construct an approximation to an appropriate sampling distribution. The m out of n bootstrap draws samples of size $m < n$, often $m \ll n$, with replacement from the original data, and can offer a similar repair for inconsistency.

Let $Y = (Y_1, \ldots, Y_n)$ be a random sample of size n from an unknown distribution F and suppose we wish to construct a confidence region for a scalar parameter $\theta \equiv \theta(F)$. The confidence set is constructed by estimating the sampling distribution of a statistic $\widehat{\theta}_n \equiv \widehat{\theta}_n(Y)$ that converges weakly to θ at a rate τ_n. Suppose also that $\widehat{\sigma}_n \equiv \widehat{\sigma}_n(Y)$ converges in probability to a constant $\sigma > 0$. Usually $\sigma^2 \equiv \sigma^2(F)$ is the asymptotic variance of $\tau_n \widehat{\theta}_n$, corresponding to Studentization, but this formulation also covers other possibilities; in particular, the un-Studentized case arises with $\widehat{\sigma}_n = 1$. Let $J_n(F)$ denote the sampling distribution of $\tau_n \widehat{\sigma}_n^{-1}(\widehat{\theta}_n - \theta)$ and suppose there exists a nondegenerate limiting distribution $J(F)$, continuous in x and such that for all real x,

$$\begin{aligned} J_n(x, F) &\equiv P_F\{\tau_n \widehat{\sigma}_n^{-1}(\widehat{\theta}_n - \theta) \leq x\} \\ &\to J(x, F) \quad \text{as } n \to \infty; \end{aligned}$$

that is, $J_n(F)$ converges weakly to $J(F)$.

Let $W_1, \ldots, W_S$ be the $S = \binom{n}{m}$ subsets of $(Y_1, \ldots, Y_n)$ of size m, and let $\widehat{\theta}_{n,m,s}$ and $\widehat{\sigma}_{n,m,s}$ be the values of $\widehat{\theta}_n$ and $\widehat{\sigma}_n$ calculated from W_s. The subsampling distribution of $\tau_n \widehat{\sigma}_n^{-1}(\widehat{\theta}_n - \theta)$, based on subsample size m, is

$$L_{n,m}(x) = S^{-1} \sum_{s=1}^{S} I\{\tau_m \widehat{\sigma}_{n,m,s}^{-1}(\widehat{\theta}_{n,m,s} - \widehat{\theta}_n) \leq x\},$$

where $I(\cdot)$ denotes the indicator function. Then it may be shown that if $m \to \infty$ and $\max(m/n, \tau_m/\tau_n) \to 0$ as $n \to \infty$, then

$$\sup_x |L_{n,m}(x) - J(x, F)| = o_p(1),$$

so that confidence sets for θ of asymptotically correct coverage can be constructed using the quantiles of $L_{n,m}$ as approximations to those of $\tau_n \widehat{\sigma}_n^{-1}(\widehat{\theta}_n - \theta)$. If S is too large for exact computation of $L_{n,m}$, a stochastic approximation is found by random sampling from $W_1, \ldots, W_S$.

If the data $(Y_1, \ldots, Y_n)$ are a time series, subsampling remains valid under a strong mixing assumption (Politis, Romano and Wolf, 1999, Appendix A), now with W_s representing the subsequence $(Y_s, Y_{s+1}, \ldots, Y_{s+m-1})$, where $s = 1, \ldots, S$ and $S = n - m + 1$.

The bootstrap approximation to $J_n(x, F)$ is $J_n(x, \widehat{F}_n)$, where $\widehat{F}_n$ is often taken to be the empirical distribution $\tilde{F}$. Bootstrap consistency and coverage results analogous to those described above have been proved for many situations in a series of articles initiated by Bickel and Freedman (1981), but under stronger conditions. Typically these results are proved by taking a metric d on the space of probability measures and showing that $d(F_n, F) \to 0$ implies weak convergence of $J_n(F_n)$ to $J(F)$. Thus the previous assumptions are strengthened so that convergence of $J_n(F)$ to $J(F)$ is locally uniform in F. Moreover, the estimator $\widehat{F}_n$ must be shown to satisfy $d(\widehat{F}_n, F) \to 0$ almost surely or in probability under F. No such extra condition is required for validity of the subsampling approach. In

counterexamples to asymptotic validity of the bootstrap, failure stems precisely from nonuniformity in the convergence.

Subsampling can often be used to remedy bootstrap inconsistency, leading to the so-called m out of n bootstrap; see Bickel, Götze and van Zwet (1997). Instead of approximating $J_n(F)$ by $J_n(\widehat{F}_n)$, we use $J_m(\widehat{F}_n)$ for some m, usually satisfying $m/n \to 0$ and $m \to \infty$ as $n \to \infty$. The resulting distribution function estimator is very similar to $L_{n,m}(x)$, the key difference being between sampling with and without replacement. The additional assumption that $m^2/n \to 0$ ensures that conclusions about asymptotic validity of the subsampling estimator are true also for the m out of n bootstrap.

Although they provide methodologies that are generally valid under minimal and easily verifiable assumptions, subsampling and the m out of n bootstrap share a number of awkward features. The most important is the optimal choice of the subsample size, which depends in a delicate way on the inference being performed. Empirical choice is often difficult, although methods that perform satisfactorily in practice have been developed (Politis, Romano and Wolf, 1999, Chapter 9).

A further theoretical cause for concern relates to the level of accuracy derived from the subsampling approach, specifically the rate of convergence of $L_{n,m}(x)$ to $J(x, F)$. This can be illustrated by letting $\widehat{\theta}_n$ represent the sample mean, for which a normal approximation to $J(x, F)$ is in error by $O_p(n^{-1/2})$. If Studentization is performed using the usual unbiased variance estimator, then the best rate of approximation achievable by the subsampling estimator $L_{n,m}(x)$ is obtained for $m \propto n^{2/3}$, but the constant of proportionality depends on the underlying population and the approximation error cannot be made smaller than $O_p(n^{-1/3})$ (Politis, Romano and Wolf, 1999, Section 10.3.2). Thus in this case the usual normal approximation is better than subsampling. Chapter 10 of Politis, Romano and Wolf (1999) described techniques by which higher-order accuracy can be squeezed from subsampling estimators; see also Bickel, Götze and van Zwet (1997), who suggested that although the m out of n bootstrap shares the same efficiency losses when a conventional bootstrap is valid, it may be more accurate than subsampling.

In summary, subsampling is valid more widely than the bootstrap and so may be regarded as superior to it in terms of first-order asymptotics. Subsampling is valid under minimal conditions both for random samples and with more complicated data structures, in particular time series, marked point processes and random fields, and it can be extended to situations where the convergence rate τ_n of the estimator is unknown. When the usual bootstrap is valid, however, its higher-order accuracy makes it preferable. Subsampling is strongly indicated only when the validity of the bootstrap is unclear but that of subsampling can be verified; see, for instance, Example 2.2.1 of Politis, Romano and Wolf (1999). When the m out of n bootstrap is valid, it seems to be preferable to subsampling, but has the same practical problems.

6. BOOTSTRAPPING SUPEREFFICIENT ESTIMATORS

More subtle considerations than those of the previous section come into play in circumstances where consistency of the conventional bootstrap depends on the true value of the parameter of interest. In important theoretical work, Beran (1997) demonstrated that in a rich class of parametric models, asymptotic superefficiency of a sequence of estimators is a sufficient condition for bootstrap failure: a conventional parametric bootstrap distribution estimator is inconsistent at any point of superefficiency, while converging to the correct limiting distribution at any other point in the parameter space. He gave a detailed analysis within the locally asymptotically normal model and cited the Hodges and Stein estimators as examples. Here we discuss the issues that arise in the context of a particular version of the Stein estimator. The lessons we learn are broader, however, because the Stein estimator is prototypical of many nonparametric smoothers.

Let $Y_1, \ldots, Y_n$ be independent k-dimensional random vectors, distributed according to the multivariate normal distribution $N_k(\theta, I)$, and define the Stein estimator by

$$T = \left(1 - \frac{k-2}{n\|\bar{Y}\|^2}\right)\bar{Y}, \quad \text{where } \bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$$

and $\|\cdot\|$ denotes the usual Euclidean norm. A conventional parametric bootstrap approximates the distribution $H(\cdot, \theta)$ of $n^{1/2}(T-\theta)$ by the conditional distribution $H(\cdot, \bar{Y})$ of $n^{1/2}(T^*-\bar{Y})$, given $\bar{Y}$; here a bootstrap sample $Y_1^*, \ldots, Y_n^*$ is a random sample from $N_k(\bar{Y}, I)$ and T^* is the Stein estimator computed from the bootstrap sample. The bootstrap estimator $H(\cdot, \bar{Y})$ is then consistent for $H(\cdot, \theta)$ for any $\theta \neq 0$, but inconsistent when $\theta = 0$; in this case, it converges to a random probability measure (Beran, 1997).

The regime considered here is one in which the dimension k is fixed and asymptotics refer to $n \to \infty$.

Beran (1995) considered an alternative regime in which the dimension $k \to \infty$. There too bootstrap asymptotics are subtle, with theoretical and numerical results differing substantially from the current regime.

In the present setting there are two natural ways to repair the bootstrap. First, we can attempt to develop diagnostics that aid us to decide whether $\theta = 0$ or not, and then use the standard estimator $H(\cdot, \bar{Y})$ only if those diagnostics indicate that it is safe to do so. Empirical diagnostics were discussed by Beran (1997) and Canty, Davison, Hinkley and Ventura (2002) and were examined critically by Samworth (2003). Second, we can use a different bootstrap procedure which yields consistency for all θ. Two approaches to this are to use the m out of n bootstrap, and to define a new estimator $\widehat{\theta}$ and approximate $H(\cdot, \theta)$ by $H(\cdot, \widehat{\theta})$, for example, taking

$$\widehat{\theta} = \begin{cases} 0, & \text{if } \|\bar{Y}\| \le cn^{-1/4}, \\ \bar{Y}, & \text{otherwise,} \end{cases} \tag{2}$$

for appropriate $c > 0$.

Both approaches lead to consistent bootstrap estimation throughout the parameter space, but both are problematic in practice. As mentioned in Section 5, the choice of m for the m out of n bootstrap is difficult but crucial, while taking $m < n$ may reduce efficiency if the ordinary bootstrap is valid. The second approach requires precise specification of the new estimator $\widehat{\theta}$ and does not guarantee improved finite sample performance at $\theta = 0$ unless c is chosen carefully. Further, both approaches can give estimators which behave poorly in neighborhoods of $\theta = 0$ compared to the conventional bootstrap (Samworth, 2003). Although these neighborhoods vanish in the limit as n increases, the practitioner is confronted with the possibility of an impaired inference compared to naive use of the standard bootstrap estimator.

As an example, consider constructing a confidence set for θ, centered at the Stein estimator. A confidence set of exact coverage $1 - \alpha$ is $\{\theta : \|T - \theta\|^2 \le w_k(\alpha, \theta)\}$, where $w_k(\alpha, \theta)$ is the upper α point of the distribution of $\|T - \theta\|^2$ under $N_k(\theta, I)$. We consider bootstrap confidence sets of nominal coverage $1 - \alpha$. Under the conventional approach these are obtained by replacing the unknown $w_k(\alpha, \theta)$ by a bootstrap estimator $w_k(\alpha, \bar{Y})$; here asymptotic correctness depends on θ. Under the modified approach, $w_k(\alpha, \theta)$ is replaced by $w_k(\alpha, \widehat{\theta})$, with $\widehat{\theta}$ defined at (2); this is always asymptotically justified.

EXAMPLE 4 (Stein estimator). We performed a simulation study for $n = 1$, 5 and 10, $k = 5$ and $\theta = \sqrt{\lambda/2} \times (-1, 1, 0, 0, 0)$ for a range of λ. For each n and λ, 50,000 confidence sets of both kinds were constructed, with bootstrap quantiles estimated by drawing 999 parametric bootstrap samples. Figure 1 displays the coverages of confidence sets of nominal coverage 95%. In the definition of the estimator $\widehat{\theta}$ we took $c = k$; we found that smaller values give much less improvement in coverage accuracy at the point of inconsistency of the conventional approach, $\theta = 0$. Although the modified bootstrap procedure has satisfactory coverage at $\theta = 0$, it performs extremely poorly near that point, giving much worse coverage accuracy than does the usual bootstrap. The coverage error of the usual bootstrap at $\theta = 0$ is less than the error of the modified procedure at neighboring values. Thus the price paid to eliminate the unsatisfactory asymptotic performance of the conventional bootstrap at $\theta = 0$ is an even greater finite sample error at other parameter values.

Both these results and similar analysis performed by Samworth (2003) for the m out of n bootstrap sug-

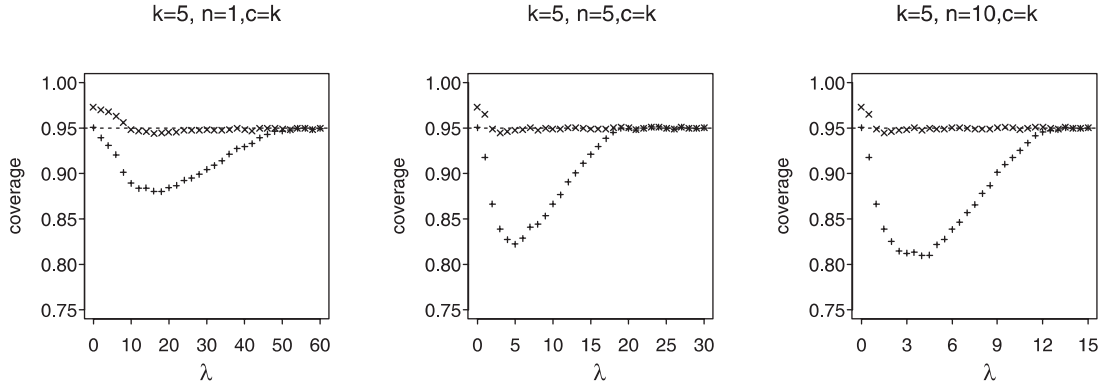

FIG. 1. *Empirical coverages of confidence sets with nominal coverage* 0.95 *in the Stein estimator example, based on* 50,000 *parametric bootstrap samples with conventional* (×) *and modified* (+) *bootstraps, with* $k = 5$ *dimensions, sample size* $n = 1, 5, 10$ *and* $c = k$.

gest that the conventional bootstrap, though asymptotically invalid for certain parameter values, might be preferable in small sample contexts to adjusted procedures that are asymptotically valid but perform inadequately in finite samples. Thus the usual bootstrap seems preferable for use in most circumstances.

7. MORE ON SIGNIFICANCE TESTS

Statistical methodology has developed to accommodate increasingly complex and large data analysis problems, and part of this methodology is significance testing. The optimistic view is that bootstrapping can always provide a solution. One area of recent interest is significance tests related to nonparametric model fits, either comparing these to parametric (or semiparametric) model fits or comparing nonparametric fits from several data sets. In many cases the asymptotic theory is both difficult and practically unhelpful, so that bootstrap resampling schemes do play an important role. Relevant references include Stute, González Manteiga and Presedo Quindimil (1998), Fan and Lin (1998), Delgado and González Manteiga (2001) and Wang and Wahba (1995). Much of the research extends the basic model-based resampling described in Chapters 6 and 7 of Davison and Hinkley (1997), for example, but it seems clear that broader theoretical development is needed to provide flexible methods.

Below we review one recent contribution by Liu and Singh (1997), the basis of which is the well-known connection between significance tests and confidence sets in parametric problems—crudely that a $1-\alpha$ confidence set for a parameter θ contains parameter values that would not be rejected in a level α significance test against an appropriate alternative. Some bootstrap procedures for calculating confidence sets were described in Section 2, where we mentioned that they can be used to avoid resampling from null hypothesis models in tests of significance.

Suppose first that $Y_1, \ldots, Y_n$ are a random sample from a distribution F, with a parameter $\theta = \theta(F)$ estimated by $\tilde{\theta}$. If we generate R conventional bootstrap samples and calculate the corresponding estimates $\tilde{\theta}_1^*, \ldots, \tilde{\theta}_R^*$, then the *empirical strength probability* (ESP) for a set null hypothesis $H_0 : \theta \in \Theta_0$ is defined to be

$$\text{ESP} = \text{proportion of } \tilde{\theta}_r^* \in \Theta_0; \tag{3}$$

it is assumed that Θ_0 has nonzero measure and a suitably smooth boundary. The ESP behaves much like a P-value, asymptotically as $n \to \infty$, if after a suitable common standardization both $\tilde{\theta}^* - \tilde{\theta}$ and $\tilde{\theta} - \theta$ have the same limiting distribution as $n \to \infty$.

One might think of the ESP as analogous to a posterior probability for Θ_0, the distribution of $\tilde{\theta}^*$ given $Y_1, \ldots, Y_n$ being an approximate posterior density. However, this is only likely to be accurate to the extent that a normal approximation for $\widehat{\theta}$ is accurate and generally is *not* second-order correct.

EXAMPLE 5 (Exponential mean). As a simple example where performance can be measured exactly, suppose that the Y_i are independent exponential variables with common mean $\theta = \mu$ and that we wish to test the null hypothesis $H_0 : \mu \le \mu_0$ versus $H_1 : \mu > \mu_0$. An exact test is possible in this problem, giving a P-value with a uniform distribution.

A bootstrap approach would use the fully efficient estimator $\tilde{\theta} = \bar{Y}$, and so

$$\text{ESP} = \Pr^*(\bar{Y}^* \le \mu_0 | Y_1, \ldots, Y_n).$$

Here $\bar{Y}^*$ is the average of a random sample of n exponential variables with mean $\bar{Y}$ if a parametric bootstrap is used, or the average of a sample of n drawn randomly with replacement from $Y_1, \ldots, Y_n$ if a nonparametric bootstrap is used. Table 4 shows how these perform for $n = 10$, 25 and 100. For the parametric bootstrap, the ESP is calculated exactly. The ESP does not work very well for small n with the parametric bootstrap, but works reasonably well for the nonparametric bootstrap.

EXAMPLE 6 (Poisson dispersion). Suppose that $Y_1, \ldots, Y_n$ is a random sample with common mean μ and variance σ^2, and that we wish to test the null hypothesis $H_0 : \theta = \mu/\sigma^2 \ge 1$ versus the alternative hypothesis $H_1 : \theta < 1$. This might be taken as a test for

TABLE 4

Percentage of times (in 500 data sets) that ESP *is less than or equal to α in one-sided tests of mean $\mu_0 = 1$ based on exponential samples of sizes $n = 10, 25$ and 100, using $R = 100$ replicates for the nonparametric bootstrap and the exact proportion for the parametric bootstrap*

		100α(%)				
n	Model	1	5	10	25	50
10	Nonparametric	1	9	14	33	59
	Parametric	5	12	18	36	60
25	Nonparametric	2	7	14	26	55
	Parametric	3	10	16	30	55
100	Nonparametric	1	7	12	28	51
	Parametric	1	8	13	28	53

TABLE 5
Percentage of times (in 500 data sets) that nonparametric ESP *is less than or equal to α in one-sided tests of dispersion, when data are a Poisson sample of size n with mean μ; $R = 100$ bootstrap replicates are used*

		$100\alpha(\%)$				
n	μ	1	5	10	25	50
50	1	2	7	13	24	59
50	10	3	9	15	27	59
10	5	4	15	24	34	68

Poisson dispersion versus overdispersion. We calculate the ESP as in (3), with $\tilde{\theta} = \bar{Y}/S^2$ the ratio of sample mean to sample variance. Then $\text{ESP} = \Pr^*(\bar{Y}^* \geq S^{*2})$. Table 5 illustrates how ESP performs for fairly large n (50) when μ is small (1) and not (10), and for small n (10) with moderate μ (5).

In the case of a point null hypothesis $H_0: \theta = \theta_0$ with two-sided alternative $H_1: \theta \neq \theta_0$, the previous approach cannot work. Instead, the ESP can be defined in terms of a confidence set for θ,

$$\text{ESP} = \max\{p : 1-p \text{ confidence set for } \theta \text{ includes } \theta_0\}. \tag{4}$$

The simplest implementation of this uses the percentile method (Section 2), in which an equitailed $(1-2\alpha)$ confidence set is the interval $(\tilde{\theta}^*_{(\alpha R)}, \tilde{\theta}^*_{((1-\alpha)R)})$. Then

$$\text{ESP} = 2\min(k, R-k)/R \quad \text{if } \tilde{\theta}^*_{(k)} \leq \theta_0 \leq \tilde{\theta}^*_{(k+1)},$$

where we define $\tilde{\theta}^*_{(0)} = -\infty$ and $\tilde{\theta}^*_{(R+1)} = +\infty$.

Of course other confidence set methods can also be used, including those based on distances. This becomes particularly relevant in more complex testing situations, such as $H_0: F \in \mathcal{F}_0$ versus $H_1: F \notin \mathcal{F}_0$, which includes the special cases $\mathcal{F}_0 = \{F : \theta(F) = \theta_0\}$ and $\mathcal{F}_0 = \{F_0\}$, a single specific distribution. If we define a distance $d(F, F_0)$ between distributions F and F_0, and use $d(\tilde{F}, F_0)$ as the test statistic for a specific F_0, then the analog of (3) is

$$\text{ESP} = \Pr^*\{\tilde{F}^* : d(\tilde{F}, \tilde{F}^*) \geq d(\tilde{F}, F_0)\};$$

this would be approximated simply by the proportion of resamples for which $d(\tilde{F}, \tilde{F}^*) \geq d(\tilde{F}, F_0)$. For general $\mathcal{F}_0$ this calculation is generalized with $d(\tilde{F}, F_0) = \min_{F \in \mathcal{F}_0} d(\tilde{F}, F)$

We can define distances in terms of empirical likelihood (Owen, 2001). Thus for a simple null hypothesis $\theta = \theta_0$ a possible test statistic is $T = L(\tilde{\theta}|\tilde{F})/L(\theta_0|\tilde{F})$, where $L(\theta)$ is the empirical likelihood. This leads to

$$\text{ESP} = \Pr^*\{\tilde{F}^* : L\{\theta(\tilde{F}^*)|\tilde{F}\} \leq L(\theta_0|\tilde{F})\},$$

which corresponds to 1 minus the smallest confidence level at which the empirical likelihood confidence set includes θ_0. Extensions to composite hypothesis problems should also be possible. In principle, the asymptotic theory developed for empirical likelihood can be used to investigate properties of ESP here.

In general, ESPs fail to behave as P-values to the extent that $\tilde{\theta}^*$ or $\tilde{F}^*$ fails to have a symmetric distribution. For this reason it seems best to use the ESP only when more specific, direct testing methods are not available for a particular problem.

8. BAGGING AND CLASSIFICATION

Since 1980 there has been an enormous amount of research on nonparametric procedures for prediction and nonparametric classification, much of it originated by computer scientists and algorithmic in approach. Some of the basic algorithms have been improved considerably using resampling as a smoothing device, which is beneficial when the basic algorithm is unstable with respect to small data perturbations. The acronym "bagging" was coined for this bootstrap aggregation. A key theoretical development in this area is the important recent discussion by Bühlmann and Yu (2002), which builds on references cited there dating back to the pioneering work by Breiman (1996a, b) and others. For data $d = \{(x_j, y_j), j = 1, \ldots, n\}$ with response y and predictor variables $x \equiv (x^{(1)}, \ldots, x^{(p)})$, imagine that a basic predictor formula $m_0(x|d_n)$ has been formed—a few examples are mentioned below. If R resampled data sets $d_1^*, \ldots, d_R^*$ are constructed and the corresponding resample predictor formulae $m_0(x|d_1^*), \ldots, m_0(x|d_R^*)$ are formed, then the empirical bagged predictor formula is

$$\hat{m}_B(x|d) = R^{-1} \sum_{r=1}^{R} m_0(x|d_r^*);$$

this is an approximation to

$$m_B(x|d) = \text{E}^*\{m_0(x|D^*)\}.$$

Typically this acts as a smoother, if smoothing is needed, and may be comparable to calculating a Bayesian posterior expectation, to the extent that the

distributions of resampled estimates of parameters in the predictor behave like posterior distributions of those parameters; see the discussion of the percentile method in Section 2. A related use of the bootstrap as an averaging device to reduce variance arises in the context of estimation, rather than reduction, of prediction and classification error (Efron, 1983, 1986; Efron and Tibshirani, 1997).

If the basic predictor m_0 is linear in the y_j, then either $m_B(x|d) = m_0(x|d)$ exactly or they are asymptotically equivalent as $n \to \infty$, under appropriate conditions on the x_j's. This happens for a linear regression formula, for example. A more practically useful variant of the linear regression formula, with screening of predictor variables, is

$$m_0(x|d) = \sum_{i=1}^{p} \widehat{\beta}_i I\big(|\widehat{\beta}_i| > c_i\big) x^{(i)},$$

where the c_i are critical levels for the estimated coefficients $\widehat{\beta}_i$ and $I(\cdot)$ is the indicator function. This is called hard thresholding. The corresponding bagged predictor,

$$m_B(x|d_n) = \sum_{i=1}^{p} \mathrm{E}^*\big\{\widehat{\beta}_i^* I\big(|\widehat{\beta}_i^*| > c_{n,i}\big)\big\} x^{(i)}, \tag{5}$$

corresponds to "soft thresholding." Bühlmann and Yu (2002) showed that for coefficients comparable in magnitude to the critical levels, bagging can reduce mean squared error for the predictor by up to 50%. Similar gains can be achieved for other unstable predictors, such as adaptive-split tree algorithms. For classification algorithms based on estimates of class membership probabilities $\widehat{\Pr}(\text{class } j|x, d)$, bagging can work by voting—that is, choosing that class which is chosen most often in R resample versions $\widehat{\Pr}(\text{class } j|x, d^*)$. Whether bagging in this context has a close parallel with Bayesian smoothing is unclear, but certainly unstable classifiers, such as tree algorithms, can be considerably improved by bagging.

A related approach known as "boosting" involves attaching weights to each datum, followed by iterative improvement of a base classifier by increasing the weights for those data that are hardest to classify with certainty. In some cases this yields dramatic reductions in classification error even relative to bagging (Freund and Schapire, 1997; Schapire, Freund, Bartlett and Lee, 1998).

9. BOOTSTRAPPING DEPENDENT DATA

It is almost self-evident that the original bootstrap is not applicable to dependent data. There has been a considerable effort to generalize the methodology to work for time series problems, but limited effort to deal with other types of stochastic process data.

Excellent surveys exist of the current state of time series bootstrap methods, including Bühlmann (2002a) and Politis (2003) in this volume. Major contributions have been various block resampling procedures, including matched-block schemes; autoregressive and other sieve schemes based on fitting models in which the number of parameters grows with the data size, but more slowly; variable length Markov chain schemes for categorical data (Bühlmann, 2002b); and nearest-neighbor resampling schemes for continuous data (Rajagopalan and Lall, 1999; Huang, 2002), which to some extent mimic variable length Markov chain schemes.

The many types of spatial data include:

1. n points $t_1, \ldots, t_n$ generated from a stochastic point process in a set $\mathcal{R}$.
2. The same as item 1 but with responses $y(t_i)$ observed at $t_i, i = 1, \ldots, n$, corresponding to a stochastic process $Y(t), t \in \mathcal{R}$.
3. Responses $y(t)$ observed on a regular lattice of points.

To maintain the relevant spatial correlation in nonparametric resampling, most research in this area discusses extensions of the block resampling idea, at least for "nice" shaped regions $\mathcal{R}$. For example, a rectangular region $\mathcal{R}$ can be partitioned into a set of b similarly shaped subrectangles (analagous to blocks), which can be randomly sampled from b times and the results pasted together to form a resample rectangle of data. Theoretical aspects of this were discussed first by Hall (1985) and most recently by Politis, Paparoditis and Romano (1999). A key difficulty is the presence of edge effects introduced by pasting independently resampled subrectangles together; no natural analogy with the matched-block or sieve approaches mentioned above has yet been identified. Lee and Lahiri (2002) discussed the application of subsampling to variogram estimation; other applications are mentioned in Section 8.3 of Davison and Hinkley (1997).

However, most of this work assumes stationarity of both the point process that generates the t's and the process for $Y(t)$ if a response y is observed as in items 2 and 3 above. This affects both the estimator

and the resampling scheme. For example, suppose that we have data of type 2 above, where the process $Y(t), t \in \mathcal{R}$, is stationary, and that we wish to estimate the mean $\mu = \int_{\mathcal{R}} Y(t)\,dt$. If the points $t_1, \ldots, t_n$ are generated from a homogeneous point process, then the estimate $\overline{Y} = n^{-1}\sum_{j=1}^{n} Y(t_j)$ is sensible and simple block resampling schemes such as outlined above may represent a good approach for large n. However, strong heterogeneity of the points $t_1, \ldots, t_n$ may make $\overline{Y}$ a poor estimate, depending on the covariance structure for $Y(t)$, and, furthermore, the relevant distribution of any estimator should condition appropriately on the point pattern—which would be an ancillary statistic in many settings—and this in turn affects the choice of an appropriate resampling scheme. At a minimum, the variable resample size n^* should be held close to n, explicitly or implicitly. Evidently this area is ripe for further research.

10. OTHER TOPICS

A common use of the bootstrap is in model selection, where it is necessary to compare empirical support for different models. About the simplest example is in straight-line regression, where the two models correspond to including the single covariate or not depending on the estimated regression coefficient. One approach to the general problem is to determine the chosen model by an estimator $\widehat{\theta}$ that falls into different regions of $\mathbb{R}^d$. That is, we suppose that $\mathbb{R}^d = \mathcal{R}_1 \cup \cdots \cup \mathcal{R}_k$ and that model i is chosen if $\widehat{\theta} \in \mathcal{R}_i$. One natural approach to assessing the uncertainty of this selection is to use the probability $\Pr^*(\widehat{\theta}^* \in \mathcal{R}_i)$ obtained in bootstrap resampling from the original data. This, however, is neither a frequentist nor a Bayesian solution: a frequentist would aim to compute a confidence level for the true parameter $\theta \in \mathcal{R}_i$ by inverting a significance test, taking $1 - \inf_{\theta \notin \mathcal{R}_i} \Pr\{u(Y, \theta) \notin \mathcal{R}_i'\}$ for some suitable pivot $u(Y, \theta)$, perhaps approximate, while a Bayesian would place a prior on θ and aim to compute the posterior probability $\Pr(\theta \in \mathcal{R}_i | Y)$. In an ingenious article, Efron and Tibshirani (1998) showed how confidence interval arguments can be modified to produce solutions to this so-called *problem of regions* that match the "ideal" solutions more closely. There seems to be a link to the discussion of empirical strength probabilities in Section 7.

The cost of computing has declined vertiginously since the bootstrap was introduced, but the amount of data available for analysis and the complexity of the procedures needing to be bootstrapped have risen equally dramatically. Thus, although computational issues matter less than they did, they remain important in some settings. One setting is iterated bootstrapping, which has become more widespread in recent years, not only for improving confidence interval algorithms as outlined in earlier sections, but also, and more fundamentally, for basic consistency checks: does the bootstrap produce reasonable solutions in *my* problem? One way to reduce the computational cost of a double bootstrap is through recycling, a version of importance sampling. Ventura (2002) showed how the original parametric simulation idea of Newton and Geyer (1994) can be adapted to the nonparametric bootstrap and gave a careful discussion of the difficulties that can arise. This and related work by Hesterberg (1999) seem to have strong links with the tilted bootstrap procedures described in Sections 3 and 4.

Recent years have seen widespread application of hierarchical and other random effects models. In parametric settings, it is often most natural to take a Bayesian point of view and to fit the parameters using Markov chain Monte Carlo algorithms, and then uncertainty analysis for both parameters and random effects is straightforwardly tackled using simulation output, at least in principle. In practice this may be unsatisfactory, either because standard parametric models fit poorly or because of concerns about the impact of the assumed priors on inferences. The specification of prior distributions can be avoided by taking a frequentist approach, under which fitting is typically performed using an expectation–maximization algorithm, possibly stochastic, and it would then seem natural to try and use the bootstrap for uncertainty analysis. Although applied in practice by Brumback and Rice (1998), Booth and Hobert (1998) and others, there seems to be no general understanding of how the bootstrap can be applied safely in this setting. McCullagh (2000) pointed out the impossibility of satisfying natural invariance conditions on sums of squares in one of the simplest random effects settings: perhaps this "smoking gun" will spur bootstrappers to take a closer look at these models. It seems possible that bootstrap methods for sample surveys, where dependence is not less real for being induced by the sampling plan, can be adapted to this setting; see the articles by Lahiri (2003) and Shao (2003) in this issue.

11. FINAL REMARKS

The simple yet powerful idea of the original bootstrap has led to an explosion of research, for reasons

that Beran (2003) mentions in this issue. That research has shown great inventiveness on several fronts, especially the theoretical. The most challenging areas are always those where current resampling technology either fails completely or gives inaccurate results. Consistency is an asymptotic property that is never strong enough for reliable data analysis, but its failure is important to discovering where new methodology needs to be developed, and an understanding of it will continue to be helpful to making theoretical progress, just as empirical procedures for assessing bootstrap success are essential tools for reliable data analysis (Canty et al., 2002).

Among the complex problems to which resampling has been applied, time series problems seem to be nearly under control because interaction between theory and methodology has paid off. Complex modelling that uses thresholding techniques needs further development, but work by Beran, Bühlmann and others has given a strong boost in this area. However, non- and semiparametric bootstrapping for spatial and hierarchical data are relatively underdeveloped.

ACKNOWLEDGMENTS

This article was improved by the helpful comments of two referees. The work was supported by the Swiss National Science Foundation.

REFERENCES

BAGGERLY, K. A. (1998). Empirical likelihood as a goodness-of-fit measure. *Biometrika* **85** 535–547.

BARNDORFF-NIELSEN, O. E. (1980). Conditionality resolutions. *Biometrika* **67** 293–310.

BARNDORFF-NIELSEN, O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **70** 343–365.

BARNDORFF-NIELSEN, O. E. (1986). Inference on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika* **73** 307–322.

BARNDORFF-NIELSEN, O. E. and COX, D. R. (1979). Edgeworth and saddle-point approximations with statistical applications (with discussion). *J. Roy. Statist. Soc. Ser. B* **41** 279–312.

BARNDORFF-NIELSEN, O. E. and COX, D. R. (1994). *Inference and Asymptotics*. Chapman and Hall, London.

BERAN R. J. (1986). Simulated power functions. *Ann. Statist.* **14** 151–173.

BERAN, R. J. (1987). Prepivoting to reduce level error of confidence sets. *Biometrika* **74** 457–468.

BERAN, R. J. (1988). Prepivoting test statistics: A bootstrap view of asymptotic refinements. *J. Amer. Statist. Assoc.* **83** 687–697.

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

BERAN, R. J. (1997). Diagnosing bootstrap success. *Ann. Inst. Statist. Math.* **49** l–24.

BERAN, R. J. (2003). The impact of the bootstrap on statistical algorithms and theory. *Statist. Sci.* **18** 175–184 (this issue).

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

BICKEL, P. J. and GHOSH, J. K. (1990). A decomposition for the likelihood ratio statistic and the Bartlett correction—a Bayesian argument. *Ann. Statist.* **18** 1070–1090.

BICKEL, P. J., GÖTZE, F. and VAN ZWET, W. R. (1997). Resampling fewer than n observations: Gains, losses, and remedies for losses. *Statist. Sinica* **7** 1–32.

BOOTH, J. G. and HOBERT, J. P. (1998). Standard errors of prediction in generalized linear mixed models. *J. Amer. Statist. Assoc.* **93** 262–272.

BRAZZALE, A. R. (2000). Practical small-sample parametric inference. Ph.D. dissertation, Dept. Mathematics, Swiss Federal Institute of Technology, Lausanne.

BREIMAN, L. (1996a). Heuristics of instability and stabilization in model selection. *Ann. Statist.* **24** 2350–2383.

BREIMAN, L. (1996b). Bagging predictors. *Machine Learning* **24** 123–140.

BRETAGNOLLE, J. (1983). Lois limites du bootstrap de certaines fonctionelles. *Ann. Inst. H. Poincaré Probab. Statist.* **19** 281–296.

BRUMBACK, B. A. and RICE, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). *J. Amer. Statist. Assoc.* **93** 961–994.

BÜHLMANN, P. (2002a). Bootstraps for time series. *Statist. Sci.* **17** 52–72.

BÜHLMANN, P. (2002b). Sieve bootstrap with variable-length Markov chains for stationary categorical time series (with discussion). *J. Amer. Statist. Assoc.* **97** 443–471.

BÜHLMANN, P. and YU, B. (2002). Analyzing bagging. *Ann. Statist.* **30** 927–961.

CANTY, A. J., DAVISON, A. C., HINKLEY, D. V. and VENTURA, V. (2002). Bootstrap diagnostics. Preprint, Institute of Mathematics, Swiss Federal Institute of Technology, Lausanne.

CARPENTER, J. (1999). Test inversion bootstrap confidence intervals. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **61** 159–172.

CORCORAN, S. A. (1998). Bartlett adjustment of empirical discrepancy statistics. *Biometrika* **85** 967–972.

COX, D. R. (1980). Local ancillarity. *Biometrika* **67** 279–286.

DAVISON, A. C. (2003). *Statistical Models.* Cambridge Univ. Press.

DAVISON, A. C. and HINKLEY, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge Univ. Press.

DELGADO, M. A. and GONZÁLEZ MANTEIGA, W. (2001). Significance testing in nonparametric regression based on the bootstrap. *Ann. Statist.* **29** 1469–1507.

DICICCIO, T. J. and EFRON, B. (1992). More accurate confidence intervals in exponential families. *Biometrika* **79** 231–245.

DICICCIO, T. J. and EFRON, B. (1996). Bootstrap confidence intervals (with discussion). *Statist. Sci.* **11** 189–228.

DiCiccio, T. J., Martin, M. A. and Stern, S. E. (2001). Simple and accurate one-sided inference from signed roots of likelihood ratios. *Canad. J. Statist.* **29** 67–76.

DiCiccio, T. J. and Romano, J. P. (1990). Nonparametric confidence limits by resampling methods and least favorable families. *Internat. Statist. Rev.* **58** 59–76.

DiCiccio, T. J. and Romano, J. P. (1995). On bootstrap procedures for second-order accurate confidence limits in parametric models. *Statist. Sinica* **5** 141–160.

Durbin, J. (1980). Approximations for densities of sufficient estimators. *Biometrika* **67** 311–333.

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

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. Amer. Statist. Assoc.* **78** 316–331.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *J. Amer. Statist. Assoc.* **81** 461–470.

Efron, B. (1987). Better bootstrap confidence intervals (with discussion). *J. Amer. Statist. Assoc.* **82** 171–200.

Efron, B. and Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information (with discussion). *Biometrika* **65** 457–487.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Chapman and Hall, New York.

Efron, B. and Tibshirani, R. J. (1997). Improvements on cross-validation: The .632+ bootstrap method. *J. Amer. Statist. Assoc.* **92** 548–560.

Efron, B. and Tibshirani, R. J. (1998). The problem of regions. *Ann. Statist.* **26** 1687–1718.

Fan, J. and Lin, S. (1998). Test of significance when data are curves. *J. Amer. Statist. Assoc.* **93** 1007–1021.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.* **55** 119–139.

Garthwaite, P. H. and Buckland, S. T. (1992). Generating Monte Carlo confidence intervals by the Robbins–Monro process. *Appl. Statist.* **41** 159–171.

Hall, P. (1985). Resampling a coverage pattern. *Stochastic Process. Appl.* **20** 231–246.

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

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer, New York.

Hall, P. and Presnell, B. (1999a). Intentionally biased bootstrap methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **61** 143–158.

Hall, P. and Presnell, B. (1999b). Biased bootstrap methods for reducing the effects of contamination. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **61** 661–680.

Hall, P. and Presnell, B. (1999c). Density estimation under constraints. *J. Comput. Graph. Statist.* **8** 259–277.

Hall, P. and Wilson, S. R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics* **47** 757–762.

Härdle, W. (1989). Resampling for inference from curves. *Bull. Inst. Internat. Statist.* **53** 53–64.

Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge Univ. Press.

Hesterberg, T. C. (1999). Bootstrap tilting confidence intervals and hypothesis tests. In *Computer Science and Statistics*: *Proc. 31st Symposium on the Interface* 389–393. Interface Foundation of North America, Inc., Fairfax Station, VA.

Hinkley, D. V. (1980). Likelihood as approximate pivotal distribution. *Biometrika* **67** 287–292.

Huang, H. (2002). Scenario generation for multivariate series data using the nearest neighbor bootstrap. Ph.D. dissertation, Dept. Decision Sciences and Engineering, Rensselaer Polytechnic Institute, Troy, New York.

Lahiri, S. N. (2003). On the impact of bootstrap in survey sampling and small-area estimation. *Statist. Sci.* **18** 199–210 (this issue).

Lawless, J. (1982). *Statistical Models and Methods for Lifetime Data.* Wiley, New York.

Lee, S. M. S. and Young, G. A. (2003). Prepivoting by weighted bootstrap iteration. *Biometrika* **90** 393–410.

Lee, Y. D. and Lahiri, S. N. (2002). Least squares variogram fitting by spatial subsampling. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 837–854.

Liu, R. Y. and Singh, K. (1997). Notions of limiting P values based on data depth and bootstrap. *J. Amer. Statist. Assoc.* **92** 266–277.

Mammen, E. (1993). Bootstrap and wild bootstrap for high-dimensional linear models. *Ann. Statist.* **21** 255–285.

Martin, M. A. (1990). On bootstrap iteration for coverage correction in confidence intervals. *J. Amer. Statist. Assoc.* **85** 1105–1118.

McCullagh, P. (2000). Resampling and exchangeable arrays. *Bernoulli* **6** 285–301.

Newton, M. A. and Geyer, C. J. (1994). Bootstrap recycling: A Monte Carlo alternative to the nested bootstrap. *J. Amer. Statist. Assoc.* **89** 905–912.

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

Owen, A. B. (2001). *Empirical Likelihood.* Chapman and Hall/CRC, Boca Raton, FL.

Politis, D. N. (2003). The impact of bootstrap methods on time series analysis. *Statist. Sci.* **18** 219–230 (this issue).

Politis, D. N., Paparoditis, E. and Romano, J. P. (1999). Resampling marked point processes. In *Multivariate Analysis, Design of Experiments, and Survey Sampling* (S. Ghosh, ed.) 163–185. Dekker, New York.

Politis, D. N., Romano, J. P. and Wolf, M. (1999). *Subsampling*. Springer, New York.

Putter, H. and van Zwet, W. R. (1996). Resampling: Consistency of substitution estimators. *Ann. Statist.* **24** 2297–2318.

Rajagopalan, B. and Lall, U. (1999). A k-nearest-neighbor simulator for daily precipitation and other weather variables. *Water Resources Res.* **35** 3089–3101.

Samworth, R. J. (2003). A note on methods of restoring consistency to the bootstrap. *Biometrika*. To appear.

Schapire, R. E., Freund, Y., Bartlett, P. and Lee, W. S. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *Ann. Statist.* **26** 1651–1686.

Severini, T. A. (2000). *Likelihood Methods in Statistics*. Clarendon, Oxford.

Shao, J. (2003). Impact of the bootstrap on sample surveys. *Statist. Sci.* **18** 191–198 (this issue).

SHAO, J. and TU, D. (1995). *The Jackknife and Bootstrap*. Springer, New York.

SINGH, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist.* **91** 1187–1195.

STUTE, W., GONZÁLEZ MANTEIGA, W. and PRESEDO QUINDIMIL, M. (1998). Bootstrap approximations in model checks for regression. *J. Amer. Statist. Assoc.* **93** 141–149.

VENTURA, V. (2002). Non-parametric bootstrap recycling. *Statist. Comput.* **12** 261–273.

WANG, Y. D. and WAHBA, G. (1995). Bootstrap confidence intervals for smoothing splines and their comparison to Bayesian confidence intervals. *J. Statist. Comput. Simulation* **51** 263–279.

WU, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis (with discussion). *Ann. Statist.* **14** 1261–1350.