Bootstrap Resampling

R. Helmers

1. Introduction

B. Efron, who invented the bootstrap in 1979, recently wrote: 'Computer-intensive methods like the bootstrap greatly extend the range of classical methods, and this is the way I believe that they will most dramatically affect 21st century statistics'. The bootstrap is a computer-intensive method for estimating the variability of statistical quantities and for setting confidence regions. The name 'bootstrap' refers to the analogy with pulling oneself up by one's own bootstraps. Efron's bootstrap is to resample the data. Given observations $X_1, \ldots, X_n$ artificial bootstrap samples are drawn with replacement from $X_1, \ldots, X_n$, putting equal probability mass $\frac{1}{n}$ at each X_i. For example, with sample size $n = 5$ and distinct observations X_1, X_2, X_3, X_4, X_5 one might obtain X_3, X_3, X_1, X_5, X_4 as bootstrap sample. In fact there are 126 distinct bootstrap samples in this case.

Bootstrap resampling often gives much better estimates than traditional statistics usually provide us with. The bootstrap can also be an effective tool in many problems of statistical inference, which otherwise would have been too complicated to handle; e.g., the construction of a confidence band in nonparametric regression, testing for the number of modes of a density, or the calibration of confidence bounds. The problem of constructing a confidence band for an unknown 'regression mean' arises, e.g., if one tries to ascertain a trend in annual series of observed (air) temperatures, possibly due to the influence of 'global warming' on such data.

In this paper I will survey recent research at CWI in the general area of bootstrap resampling methods. At the same time research in this area, which takes place at Leiden University, will also be briefly reviewed. Resampling is one of the four selected areas of research in the focus area 'Computationally Intensive Methods in Stochastics' (1993-1998) of NWO. This topic was also the central theme of a two-month research workshop presented (in the summer of 1995) at the Institute of Technology, Bandung, as part of a cooperation project 'Applied Mathematics and Computational Methods' (1995-1999) between The Netherlands and Indonesia, in which CWI is one of the Dutch cooperating Institutes.

2. EFRON'S NONPARAMETRIC BOOTSTRAP

To begin with I describe Efron's nonparametric bootstrap in a simple setting and address briefly the important question: when does Efron's bootstrap work and when does it fail?

2.1. Description of the bootstrap

Suppose $X_1, \ldots, X_n$ is a random sample of size n from a population with unknown distribution function F on the real line. Let, in addition,

$$\theta = \theta(F) \tag{2.1}$$

denote a real-valued parameter which we want to estimate.

Let $T_n = T_n(X_1, \ldots, X_n)$ denote an estimator of θ, based on the data $X_1, \ldots, X_n$. Our object of interest is the distribution of $n^{\frac{1}{2}}(T_n - \theta)$, i.e., we define

$$G_n(x) = P(n^{\frac{1}{2}}(T_n - \theta) \leq x), \ -\infty < x < \infty, \tag{2.2}$$

where P denotes 'probability' corresponding to F. Clearly G_n, the exact distribution of $n^{\frac{1}{2}}(T_n - \theta)$, is unknown, because F is not known to us, but we can try to estimate it. The Efron's nonparametric bootstrap estimator (approximation) of G_n is given by

$$G_n^*(x) = P_n^*(n^{\frac{1}{2}}(T_n^* - \theta_n) \leq x), \ -\infty < x < \infty. \tag{2.3}$$

Here $T_n^* = T_n(X_1^*, \ldots, X_n^*)$, where $X_1^*, \ldots, X_n^*$ denotes an artifical random sample—the bootstrap sample—from $\hat{F}_n$, the empirical distribution function of the original observations $X_1, \ldots, X_n$, and $\theta_n = \theta(\hat{F}_n)$. Note that $\hat{F}_n$ is the random distribution—a step function—which puts probability mass $\frac{1}{n}$ at each of the X_i's $(1 \leq i \leq n)$, sometimes referred to as the resampling distribution. The empirical distribution function $\hat{F}_n$ is illustrated in figure 1. Finally, P_n^* denotes 'probability' corresponding to $\hat{F}_n$, conditionally given $\hat{F}_n$, i.e., given the observations $X_1, \ldots, X_n$. To emphasize the fact that G_n^* is a conditional distribution, one may as well write

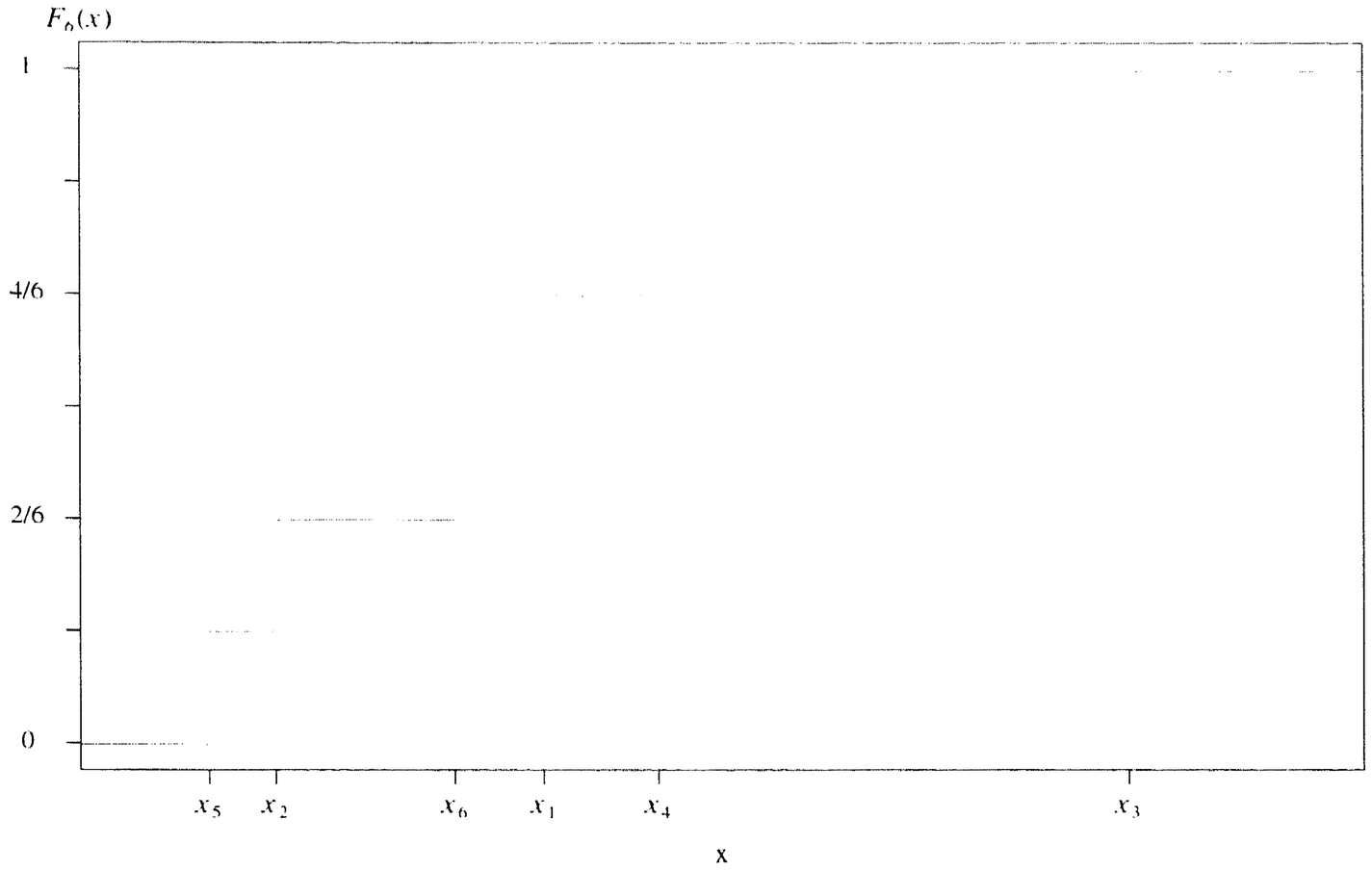

Figure 1. Empirical distribution function based on observations $x_1, \ldots, x_6$.

$$G_n^*(x) = P_n^*(n^{\frac{1}{2}}(T_n^* - \theta_n) \leq x | X_1, \ldots, X_n), \ -\infty < x < \infty, \qquad (2.4)$$

instead of (2.3). Obviously, given the observed values $X_1, \ldots, X_n$ in our sample, $\hat{F}_n$ is completely known, and—at least in principle—G_n^* is also completely known. We may view G_n^* as the empirical counterpart in the 'bootstrap world' to G_n in the 'real world'. In practice, exact computation of G_n^* by complete enumeration is usually impossible (even in our sophisticated computer age): for a sample $X_1, \ldots, X_n$ of n distinct numbers there are $\binom{2n-1}{n}$ distinct bootstrap samples. For $n = 10$ already near to 100,000 bootstrap samples have to be enumerated, so very soon this method becomes unfeasible and we have to turn to another solution: Monte-Carlo simulation.

In a sense, this boils down to repeatedly drawing a random bootstrap sample from all possible bootstrap samples. We fix a large number B. With the use of the computer, we generate a bootstrap sample and calculate the resulting value of $n^{\frac{1}{2}}(T_n^* - \theta_n)$. By repeating this procedure B times, we obtain B values, say $n^{\frac{1}{2}}(T_{n,1}^* - \theta_n), \ldots, n^{\frac{1}{2}}(T_{n,B}^* - \theta_n)$, which give an accurate Monte-Carlo estimate to the theoretical bootstrap distribution G_n^* of $n^{\frac{1}{2}}(T_n^* - \theta_n)$. Monte-Carlo simulation was of course already well established before the invention of the bootstrap, but it finds a very natural place here. Generating a bootstrap sample amounts to randomly drawing a sample of size n with replacement from $X_1, \ldots, X_n$. The Monte-Carlo procedure introduces a second source of randomness. However, by choosing

B suitable large we can control the Monte-Carlo error and make sure that it is negligible in comparison with the bootstrap approximation error.

2.2. Operation of the bootstrap

When does Efron's bootstrap work? The consistency of the bootstrap approximation G_n^*, viewed as an estimate of G_n, i.e., we require

$$\sup_x |G_n(x) - G_n^*(x)| \to 0, \text{ as } n \to \infty \tag{2.5}$$

to hold, with P-probability one (i.e., for almost all sequences $X_1, X_2, \ldots$), or a slightly weaker version of it, namely that (2.5) holds only in P-probability, rather than P-almost surely, is generally viewed as an absolute prerequisite for Efron's bootstrap to work in the problem at hand. Of course, the assertion (2.5) is only a first order asymptotic result, and the error committed, when the bootstrap is applied in finite samples—say, with sample size $n = 20$—may still be quite large.

In the important special case that $\theta(F) = \mu = \int x dF(x)$, the population mean, and $T_n = \bar{X}_n = n^{-1} \sum_{i=1}^n X_i$, the sample mean, a by now classical result asserts that (2.5) holds true, i.e., Efron's bootstrap works, provided the variance σ^2 of the underlying distribution F is finite. If σ^2 is infinite the situation becomes more complex: it has been proved that Efron's bootstrap still works, provided F is in the domain of attraction of the normal law; otherwise Efron's bootstrap fails.

Bootstrap resampling can also be used to estimate functionals of G_n, e.g., its variance, rather then G_n itself. W.R. van Zwet (1994) has recently studied the performance of Efron's bootstrap estimate of variance for arbitrary symmetric statistics $T_n = T_n(X_1, \ldots, X_n)$ with finite second moment using the Hoeffding decomposition. He showed that Efron's bootstrap will typically work, provided $\sum_{i=1}^n E(T_n|X_i)$, the linear term in the Hoeffding decomposition of T_n, is the dominant one and the higher order terms in the Hoeffding decomposition tend to zero rather fast. The requirement concerning the linear term is also shown to be a necessary condition for the consistency of Efron's bootstrap; otherwise (2.5) generally fails to hold. A specific case for which Efron's bootstrap works—namely Serfling's class of generalized L-statistics—is investigated by R. Helmers, P. Janssen, and R. Serfling (1990).

H. Putter and Van Zwet (1993) (c.f. also chapter 2 of the Ph.D. thesis of Putter (1994)) emphasized the importance of a proper choice of the resampling distribution (not necessarily the empirical distribution $\hat{F}_n$, as in Efron's nonparametric bootstrap). Let $\tau_n(F)$ denote the distribution of a statistical quantity $R_n = R_n(X_1, \ldots, X_n; F)$. If $\tilde{F}_n$ denotes the resampling distribution, $\tilde{F}_n = \tilde{F}_n(X_1, \ldots, X_n)$ being an estimate of F, then the bootstrap estimate of $\tau_n(F)$ becomes $\tau_n(\tilde{F}_n)$. Note that $\tilde{F}_n$ may be very different from the empirical distribution $\hat{F}_n$, e.g., one may consider $\tilde{F}_n = F_{\hat{\theta}_n}$, when

it is a priori known that F belongs to a parametric model $\{F_\theta, \theta \in \Theta\}$, the finite-dimensional parameter θ is estimated by $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$, a consistent estimator of θ (parametric bootstrap). Putter and Van Zwet (1993) have recently proved a general result concerning the consistency of bootstrap estimates, with general resampling distribution $\tilde{F}_n$.

3. ACCURACY OF BOOTSTRAP ESTIMATES

3.1. Smooth cases

In the previous section we have seen that Efron's bootstrap is consistent for the case of the sample mean $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$, provided the underlying distribution F of the observations has a finite second moment. With

$$G_n(x) = P(n^{\frac{1}{2}}(\bar{X}_n - \mu) \leq x) \tag{3.1}$$

and

$$G_n^*(x) = P_n^*(n^{\frac{1}{2}}(\bar{X}_n^* - \bar{X}_n) \leq x) \tag{3.2}$$

we have, with P-probability 1,

$$\sup_x |G_n(x) - G_n^*(x)| \to 0, \text{ as } n \to \infty \tag{3.3}$$

whenever $0 < \int x^2 dF(x) < \infty$. However, the question remains: how well does Efron's bootstrap estimate G_n^* approximates G_n? The answer is that typically the rate of convergence in (3.3) is of the classical order $n^{-\frac{1}{2}}$. The famous Berry-Esseen theorem asserts that the accuracy of the normal approximation is of the same order $n^{-\frac{1}{2}}$, provided $\int |x|^3 dF(x) < \infty$. However, we can easily improve the accuracy of our bootstrap estimate, by first employing 'Studentization'. That is, instead of the statistical quantity $n^{\frac{1}{2}}(\bar{X}_n - \mu)$ and its bootstrap version $n^{\frac{1}{2}}(\bar{X}_n^* - \bar{X}_n)$, we consider the old and famous Student t statistic $n^{\frac{1}{2}}(\bar{X}_n - \mu)/S_n$ and its bootstrap counterpart $n^{\frac{1}{2}}(\bar{X}_n^* - \bar{X}_n)/S_n^*$, with respective distribution functions

$$G_{ns}(x) = P(n^{\frac{1}{2}}(\bar{X}_n - \mu)/S_n \leq x), \ -\infty < x < \infty, \tag{3.4}$$

and

$$G_{ns}^*(x) = P_n^*(n^{\frac{1}{2}}(\bar{X}_n^* - \bar{X}_n)/S_n^* \leq x), \ -\infty < x < \infty. \tag{3.5}$$

where $S_n^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ denotes the sample variance. Note that S_n^{*2} is nothing but S_n^2, with the X_i's replaced by the X_i^*'s. We note in passing that, if F is normal, G_{ns} of course reduces to the well-known Student t distribution with $n-1$ degrees of freedom. In general, however, the exact distribution G_{ns} of Student's t is unknown, but we can try to estimate it, e.g., by using the bootstrap. Similarly, as in (3.3) we have that

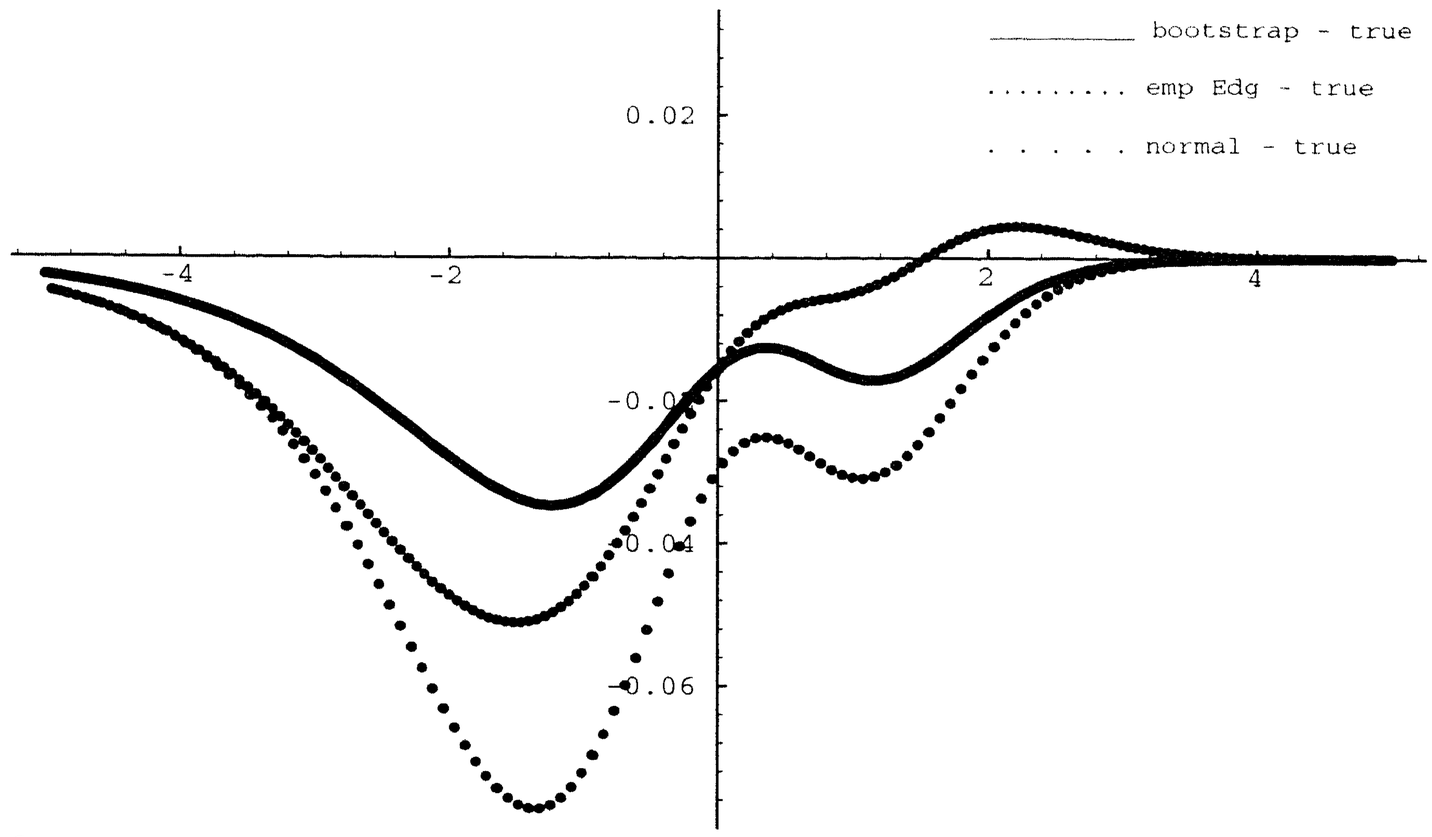

Figure 2. Three approximations; $n = 20$; F exponential.

$$\sup_x |G_{ns}(x) - G^*_{ns}(x)| \to 0, \text{ as } n \to \infty, \tag{3.6}$$

but now the rate of convergence is faster: in fact, one can show that in P-probability

$$n^{\frac{1}{2}} \sup_x |G_{ns}(x) - G^*_{ns}(x)| \to 0, \text{ as } n \to \infty \tag{3.7}$$

under rather weak conditions. Under somewhat more stringent assumptions, one can prove that $\sup_x |G_{ns}(x) - G^*_{ns}(x)|$, the accuracy of the bootstrap approximation, is of the exact order n^{-1} in P-probability. In contrast, the normal approximation for Student's t possesses the classical Berry-Esseen type error of order $n^{-\frac{1}{2}}$.

250

In other words: the bootstrap estimate G^*_{ns} is asymptotically closer to G_{ns} than the standard normal distribution. This 'bootstrap is better than normal' property of Efron's bootstrap for the Student t statistic clearly suggests the benificial effect of 'Studentization' before bootstrapping for this important special case. A Monte-Carlo result, which supports this claim, is presented in figure 2 (borrowed from Putter's thesis (1994)). We consider the special case that F is exponential. First of all, the distribution G_{ns} was approximated by Monte-Carlo using 10^7 samples. Next a sample of size $n = 20$ was drawn from a standard exponential distribution and—based on this sample—the distribution G_{ns} was estimated in three ways, first using the classical normal approximation, secondly using the bootstrap G^*_{ns} (as in (3.5), using Monte-Carlo simulation with $B = 10^6$). With this choice

of B, we are pretty certain that the Monte-Carlo error is negligible. Note that we should take care that too low a choice for B doesn't ruin the second order accuracy of the bootstrap estimate G^*_{ns}. In fact, it is easily checked that this means that the Monte-Carlo error —which is of order $\frac{1}{\sqrt{B}}$— should be of a smaller order than $\frac{1}{n}$, the accuracy of the bootstrap approximation, i.e., B should be of a larger order than n^2. The third way uses empirical Edgeworth expansion (EEE). To make the differences between these three methods discernible we have plotted for each of the three methods the resulting estimate minus the target distribution G_{ns}. The graph that lies closest to zero corresponds therefore to the best approximation. It is clearly seen that both bootstrap and Edgeworth expansion outperform the normal approximation. The bootstrap performs slightly better than EEE, due to the fact that the bootstrap also implicitly estimates higher order terms in the Edgeworth expansion consistently.

3.2. Non-smooth cases

The above result for Student's t is in fact already known for about 10 years (cf., e.g., the references in [2]) and it is generally viewed as an important argument in favour of Efron's bootstrap. Helmers proved (1991) that the 'bootstrap is better than normal' property also holds true for more complicated nonlinear statistics like Hoeffding's famous class of U-statistics. The extension of the 'bootstrap is better than normal' property to arbitrary Studentized symmetric statistics is still an interesting open problem at present. In any case, however, the quadratic and higher order terms in the Hoeffding decomposition for a symmetric statistic $T_n = T_n(X_1, \ldots, X_n)$ should be of a required order of magnitude, otherwise the speed of bootstrap convergence asserted in (3.7) fails to hold. An important specific example of the latter is the case of the median, and more generally, quantiles. In such 'non-smooth' cases (the parameter of interest, e.g., $\theta = \theta(F) = F^{-1}(\frac{1}{2})$ is a much less smooth functional of F, then the parameter $\theta = \theta(F) = \int x dF(x)$) we have a much slower rate (roughly of order $n^{-\frac{1}{4}}$) of convergence of Efron's bootstrap approximation. In fact, although Efron's bootstrap for the median is consistent, it is worthless in practice, even for a sample size n as large as 100. In the computer calculations that led to figure 3 we have generated a sample of size $n = 100$ from a standard normal distribution. As a result we find that the difference with the true distribution function is maximized at $x = 0.39$; at this point the true distribution equals 0.657 while the bootstrap approximation yields 0.935, which means a relative error of more than 40%. It is well known that the smoothed bootstrap, where resampling is done from a smoothed version of the empirical distribution, results in a better approximation. For the smoothed bootstrap we used a normal kernel and a bandwidth $h = 0.1$, and indeed the smoothed bootstrap seems to perform much better. A more sophisticated choice of kernel and bandwidth will

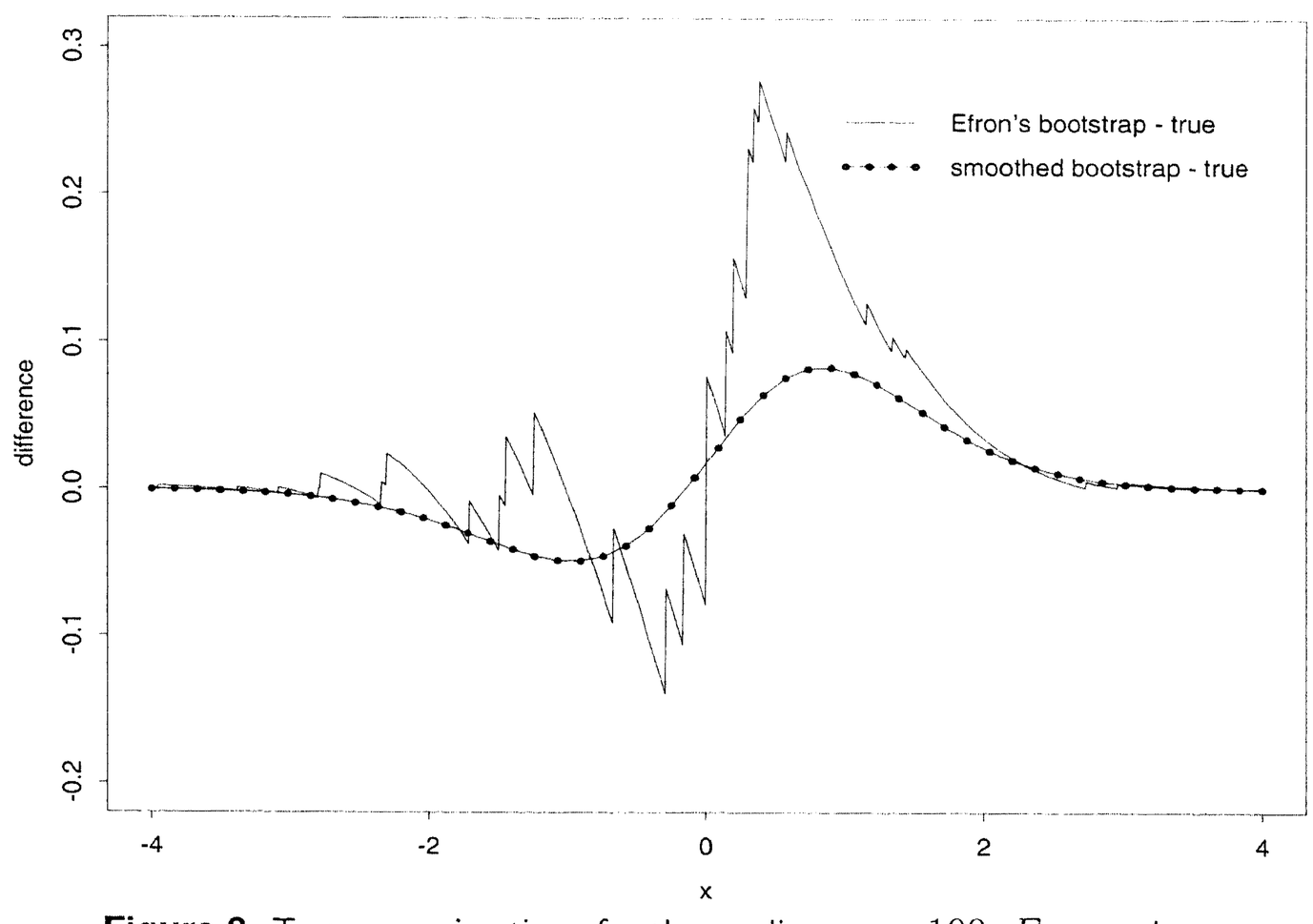

Figure 3. Two approximations for the median; $n = 100$; F normal.

presumably lead to a further reduction of the error of the smooth bootstrap approximation.

Related work for the more general case of U-quantiles can be found in a contribution of Helmers, Janssen and N. Veraverbeke to [4] (cf. Helmers, M. Hušková (1994) for an extension to multivariate U-quantiles). Specific examples of interest of U-quantiles are the well-known Hodges-Lehmann estimator of location, which is given by the median of all pairwise averages, and an estimator of spread proposed by Bickel and Lehmann. Young, p. 392 in a prominent recent review paper [5], feels that 'this sort of work is important', because 'the contexts to which the results apply are highly relevant to precisely the sort of circumstances—when there is limited knowledge about the underlying distribution—for which bootstrap was designed'.

4. Applications

To conclude I briefly discuss two selected topics of current interest in bootstrap theory and its applications: resampling methods for finite populations, and spatial bootstrapping.

Resampling methods for finite populations is an important topic of current interest. Helmers and M.H. Wegkamp considered (1995) the situation where the finite population is viewed as a realization of a certain superpopulation model (heteroscedastic linear regression, without intercept). This

Figure 4. Observed oilspot.

enables us to incorporate auxiliary information (past experience) in the statistical analysis. The authors first came across this problem in a 1994 statistical consultation project at CWI with The Netherlands postal services PTT Post. In this setup a new resampling scheme called 'two-stage wild bootstrapping' is proposed and studied. The basic probabilistic tool we employ in our mathematical analysis is the celebrated Erdös-Rényi central limit theorem for samples without replacement from a finite population.

Bootstrapping with spatial data is very clearly an important area for future work in the research group 'Image analysis and spatial stochastics' of CWI. We briefly describe here a practical application in which spatial bootstrapping is used. In a project commissioned by the North Sea Directorate, Ministry of Public Works the problem is to estimate the intensity of oilpollution in the North Sea. The available real data sets ('marked planar point patterns') consist of the locations and sizes (marks) of the oilspots observed (cf. figure 4) by a surveillance aircraft. A planar inhomogeneous Poisson point process with intensity function $\lambda(\cdot, \theta)$—parameterized by a finite-dimensional parameter θ—was used as a spatial (parametric) model for the locations of (the centres of) oilspots. The parameterization enables one to incorporate the available a priori knowledge about oil pollution, such as the location of sources of oilpollution (i.e. shipping areas or off-shore locations) and the intensity of shipping in various regions. However, nothing

seems to be known about the distribution of the volumes (marks) of oilspots, but we can of course use the sizes of the observed oilspots to estimate it (nonparametric approach). In this setup a simple semiparametric form of spatial bootstrapping was developed in order to estimate the accuracy of the estimated total amount of oilpollution in the North Sea.

5. ACKNOWLEDGEMENT

I want to thank H. Putter for his contributions to the present paper.

The interested reader is referred to [1] for an excellent introduction to the bootstrap. Uses of Edgeworth expansions in the mathematical analysis of Efron's bootstrap is the topic of the research monograph [2]. Additional information on the bootstrap may also be found in the proceedings volume [4] and discussion paper [5]. The present article is basically a shortened non-technical revision of [3]. The latter reference also contains a more complete list of references.

REFERENCES

1. B. EFRON, R.J. TIBSHIRANI (1993). *An Introduction to the Bootstrap*, Chapman and Hall, New York.
2. P. HALL (1992). *The Bootstrap and Edgeworth expansion*, Springer, New York.
3. R. HELMERS, H. PUTTER (1995). Bootstrap resampling: a survey of recent research in The Netherlands, CWI Report BS-R9517. To appear in *Proceedings of the SEAMS Regional Conference on Mathematical Analysis and Statistics*, Yogyakarta, Indonesia, July 10-13, 1995.
4. R. HELMERS, P. JANSSEN, N. VERAVERBEKE (1992). Bootstrapping U-quantiles. R. LEPAGE, L. BILLARD (eds.). *Exploring the Limits of Bootstrap*, Wiley, New York.
5. G.A. YOUNG (1994). Bootstrap: more than a stab in the dark? *Statistical Science 9(3)*, 382-415.