Journal of the Korean Statistical Society (2003), **32**: *3, pp 213–224*

INVITED PAPER

UNORTHODOX BOOTSTRAPS[†]

PETER J. BICKEL[1]

ABSTRACT

We give an overview of results which have appeared or will appear elsewhere demonstrating that by suitably modifying the bootstrap principle, its applicability can be greatly enhanced. Although we state our results for the *iid* case, extensions are, at least heuristically, easy.

AMS 2000 subject classifications. Primary 62G09.
Keywords. Subsampling, m out of n bootstrap, nonparametric bootstrap, confidence band bootstrap.

1. INTRODUCTION

Since its introduction in 1979, Efron's nonparametric bootstrap has proved its value in an ever increasing circle of applications. From the beginning, extensions to regressions models were presented. These have been developed in many directions and the original principle has further been extended to time series and other dependent data problems. Nevertheless, there has also accumulated a body of evidence reviewed and added to in Mammen (1992) and Bickel *et al.* (1997) which clearly indicates weaknesses of the original, weaknesses which are due to its "unregularized" nature.

In this paper, we shall review a general approach for regularizing the bootstrap discovered by Politis and Romano (1994) and independently by Götze (1993) and developed by Bickel and Sakov (2002) and Götze and Rauckauskas (2002) as well as a quite different modification of Efron's method, implicitly suggested by Beran (1986) and developed by Bickel and Ren (2002). In Section 2, we give a series of examples of nonparametric bootstrap failure and the m out of n bootstraps. In Section 3, we discuss a proposal for choice of m and its application in setting

[†] Supported in part by NSF 0104075.

[1] Department of Statistics, University of California, Berkeley, U.S.A.

confidence bounds for extrema. Finally, in Section 4, we review the Beran-Bickel-Ren approach to setting critical values in testing. We do not give any proofs, referring to the existing literature as appropriate.

2. Failure of the Efron's Bootstrap and Subsampling without and with Replacement

Our setting throughout this paper will be that of observing $X_1, \dots, X_n$ which are independent and identically distributed as $X \in \mathcal{X}$, $X \sim P \in \mathcal{P} \subset \mathcal{M}$, where $\mathcal{M}$ is the collection of all probability distributions on $\mathcal{X}$ with respect to an appropriate σ field which we suppress. Let

$$P_n \equiv \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

denote the empirical distribution of the samples placing mass n^{-1} at each X_i. Efron's proposal can then be viewed abstractly as follows.

Suppose $\tilde{\mathcal{M}} \supset \mathcal{P} \cup \{\text{All distributions with finite support}\}$. We are given functions $T_n : \tilde{\mathcal{M}} \times \mathcal{P} \to \mathcal{T}$ such that $T_n(P_n, P) \equiv T_n(X_1, \dots, X_n, P)$ is a random element (measurable with respect to the specified σ field on $\mathcal{X}^n$ and a σ field on $\mathcal{T}$). Suppose

$$\mathcal{L}_P(T_n(P_n, P)) \Longrightarrow \mathcal{L}_P$$

where $\mathcal{L}_P(\cdot)$ denotes law and $\Longrightarrow$ is weak convergence in the sense of Hoffman-Jorgensen with $\mathcal{L}_P$ concentrating on a σ compact metric subspace of $\mathcal{T}$. We refer to van der Vaart and Wellner (1996) for the technical meaning of these statements. As we illustrate in the examples that follow, we are really interested in real valued parameters $\nu(\mathcal{L}_P(T_n(P_n, P))) \equiv \theta_n(P)$ which converge to parameters $\nu(\mathcal{L}_P)$ on $\mathcal{P}$ and are such that $\theta_n(P_n)$ can be defined on $\tilde{\mathcal{M}}$. The Efron's prescription is then to use the "plug-in estimates" $\theta_n(P_n)$ to estimate $\theta_n(P)$ or equivalently $\theta(P) \equiv \nu(\mathcal{L}_P)$. Note that $\theta_n(P_n)$ is ν applied to the distribution of $T_n(P_n^*, P_n) \equiv T_n(X_1^*, \dots, X_n^*, P_n)$ where $X_1^*, \dots, X_n^*$ are drawn *iid* from P_n (given $X_1, \dots, X_n$). Since drawing an *iid* sample from P_n is equivalent to sampling from the original sample with replacement, the bootstrap is often referred to as "resampling". Implementing the bootstrap typically involves Monte Carlo. That is, $\nu(\mathcal{L}_{P_n}(T_n(X_1^*, \dots, X_n^*, P_n)))$ is approximated by drawing *iid* samples $(X_{1b}^*, \dots, X_{nb}^*)$, $1 \le b \le B$ from P_n (given $X_1, \dots, X_n$) and actually estimating $\theta_n(P)$ by using ν applied to the empirical distribution of $T_n(X_{1b}^*, \dots, X_{nb}^*, P_n)$,

$1 \leq b \leq B$. The bootstrap is said *to work* if $\theta_n(P_n) \xrightarrow{P} \theta(P)$ in some uniform fashion for $\mathcal{P}$.

2.1. Usual examples - Functions

U1 : *Nothing novel.* If $T_n(P_n, P) \equiv \mu(P)$, for instance, $\mathcal{X} = \mathbb{R}$, $\mu(P) = \int x\ dP$ and $\mathcal{P} = \{P : \int x^2 dP(x) < \infty\}$, then $T_n(P_n^*, P_n) = \mu(P_n)$, the plug-in estimate, and $\mu(P_n) = \bar{X}$ in the above case. In words, the bootstrap in such situations is just the usual NPMLE, the nonparametric maximum likelihood estimate.

U2 : $T_n(P_n, P) = r_n(\mu(P_n) - \mu(P))$. If $\mu(P) = \int x dP$ and $\mathcal{P}$ as above, then

$$T_n(P_n, P) = n^{1/2}(\bar{X} - \mu(P)) \ .$$

If $\mu(P) = F^{-1}(1/2)$, $F(x) \equiv P[X \leq x]$ and $\mathcal{P} = \{F : F$ is a distribution on $\mathbb{R}$ and has a finite positive derivative f at $F^{-1}(1/2)\}$, then

$$T_n(P_n, P) = n^{1/2}\left\{\text{med}(X_1, \ldots, X_n) - F^{-1}(\frac{1}{2})\right\} \ .$$

2.2. Usual examples - Parameters of $\mathcal{L}_P$

U3 : $\theta_n(P) = r_n^2 \text{Var}_P\{\mu(P_n)\}$. The classical example is $n\text{Var}_P\{\text{med}(X_1, \ldots, X_n)\}$. Note that in this case, although $\theta_n(P_n)$ is well defined for all $P \in \tilde{\mathcal{M}}$,

$$\theta_n(P) \to \theta(P) = \frac{1}{4f^2(F^{-1}(1/2))}$$

only on $\mathcal{P}$.

U4 : $\theta_n(P)$ is a quantile. For example,

$$P\left\{n^{1/2}(\mu(P_n) - \mu(P)) \leq \theta_n(P, \alpha)\right\} = 1 - \alpha \ .$$

So, $\theta_n(P, \alpha)$ is the $(1 - \alpha)$ quantile of $\mathcal{L}_P(T_n(P_n, P))$. Then, $\mu(P_n) - n^{-1/2}\theta_n(P_n, \alpha)$ is used as an asymptotic $1 - \alpha$ lower confidence bound for $\mu(P)$.

Less obvious is the possible use of Efron's percentile bound, the lower α quantile of $\mu(P_n^*)$ itself as an asymptotic $(1 - \alpha)$ LCB. A discussion of successful application of the technique in such examples as well as bias estimation and other uses, is given in Efron and Tibshirani (1993), Hall (1992) and Bickel and Doksum (2004), for instance.

2.3. Counter examples

It is not surprising that $\theta_n(P_n)$ can misbehave badly since $\theta(P_n)$, which as we noted is just the nonparametric maximum likelihood estimate, is known to do so, for instance, for the family of star shaped distributions (Barlow *et al.*, 1972).

F1 : The simplest example is estimation of a density. Let $\mathcal{X} = \mathbb{R}$ and $\mathcal{P} = \{P : P \text{ has density } f \text{ at } c\}$. Let

$$\theta_n(P) = \frac{n}{2}\left\{F(c+n^{-1}) - F(c-n^{-1})\right\} .$$

Then,

$$\theta_n(P) \to f(c) \qquad \text{on } \mathcal{P}$$

but $\theta_n(P_n) \sim (1/2)\text{Bin}\,(n, \theta_n(P)/n)$ where Bin denotes the binomial distribution. Finally, $\theta_n(P_n) \Longrightarrow (1/2)\mathcal{P}(2f(c))$ where $\mathcal{P}$ denotes a Poisson variable and $\Longrightarrow$ indicates weak convergence of the distribution with probability 1. The bootstrap evidently doesn't work! If we replace $c \pm n^{-1}$ by $c \pm n^{-2}$ say, in $\theta_n(P)$, then it is easy to see that $\theta_n(P_n)$ converges to a delta function.

F2 : A more interesting example of the same type is presented in Bickel and Freedman (1981) in the setting of a confidence bound on the upper endpoint of a distribution with support bounded above where we are using $\max(X_1, \ldots, X_n)$ as an estimate. The general failure of the bootstrap for confidence bounds for extrema is discussed in Athreya and Fukuchi (1994).

F3 : That the nonparametric bootstrap fails in the naive setting of critical values for test statistics was in principle observed already by Freedman (1981) and Beran (1986). For instance, consider testing $H : \mu(P) = 0$ *vs.* $K : \mu(P) > 0$ with $\mu(P) = E_P X$ using $\sqrt{n}\bar{X}$. Then, the $(1-\alpha)$ quantile of the (bootstrap) distribution of $\sqrt{n}\bar{X}^*$ does not converge to the appropriate Gaussian quantile. The problem is that, since the bootstrap distribution of $\sqrt{n}(\bar{X}^* - \bar{X})$ behaves appropriately, that of $\sqrt{n}\bar{X}^* = \sqrt{n}(\bar{X}^* - \bar{X}) + \sqrt{n}\bar{X}$ cannot.

For more on problems of all of these, see Mammen (1992) and Bickel *et al.* (1997).

The essential difficulty lies in the "irregularity" of the parameters $\theta(P)$ which $\theta_n(P)$ converges to. Politis and Romano (1994) and independently Götze (1993) and Bickel *et al.* (1997) noted that regularization of this procedure was possible quite generally.

The first of our nonorthodox types of bootstraps are the m out of n with (WR) and without replacement (WOR) bootstraps. Informally, we view as our goal estimating $\theta(P)$. We proceed by approximating $\theta(P)$ not by $\theta_n(P)$ but rather by $\theta_{m(n)}(P)$ where $m(n) \to \infty$ but $m(n)/n \to 0$, and then consider estimating this parameter by resampling from $X_1, \ldots, X_n$ with and without replacement. Here are formal definitions and the statement of two theorems taken from Bickel *et al.* (1997). We refer to that paper for proofs.

Let h be a bounded real valued function defined on the range of T_n, for instance, $t \to 1(t \le t_0)$. We view as our goal estimation of $\theta_n(P) \equiv E_P(h(T_n(P_n, P)))$. More complicated parameters ν such as quantiles are governed by the same heuristics and results as those we detail below. Here are the procedures as detailed in Bickel *et al.* (1997),

i) *The n/n bootstrap* (The nonparametric bootstrap). Let

$$\begin{aligned} B_n(P) &= E^* h(T_n(P_n^*, P)) \\ &= n^{-n} \sum_{(i_1,\ldots,i_n)} h(T_n(X_{i_1}, \ldots, X_{i_n}, P)). \end{aligned}$$

Then, $B_n \equiv B_n(P_n) = \theta_n(P_n)$ is the n/n bootstrap.

ii) *The m/n bootstrap.* Let

$$B_{m,n}(P) \equiv n^{-m} \sum_{(i_1,\ldots,i_m)} h(T_m(X_{i_1}, \ldots, X_{i_m}, P)).$$

Then, $B_{m,n} \equiv B_{m,n}(P_n) = \theta_m(P_n)$ is the m/n bootstrap.

iii) *The $\binom{n}{m}$ bootstrap.* Let

$$J_{m,n}(P) = \binom{n}{m}^{-1} \sum_{i_1 < \cdots < i_m} h(T_m(X_{i_1}, \ldots, X_{i_m}, P)).$$

Then, $J_{m,n} \equiv J_{m,n}(P_n)$ is the $\binom{n}{m}$ bootstrap.

THEOREM 2.1. *Suppose $m/n \to 0$, $m \to \infty$. Then,*

$$J_{m,n}(P) = \theta_m(P) + O_p\left\{(m/n)^{1/2}\right\}. \tag{2.1}$$

If h is continuous and

$$T_m(X_1,\ldots,X_m,P) = T_m(X_1,\ldots,X_m,P_n) + o_p(1), \tag{2.2}$$

then

$$J_{m,n} = \theta_m(P) + o_p(1). \tag{2.3}$$

Let $\mathbf{X}_j^{(i)} = (X_j,\ldots,X_j)_{1\times i}$ and

$$h_{i_1,\ldots,i_r}(X_1,\ldots,X_r) = \frac{1}{r!} \sum_{1\le j_1\ne\cdots\ne j_r\le r} h(T_m(\mathbf{X}_{j_1}^{(i_1)},\ldots,\mathbf{X}_{j_r}^{(i_r)},P)). \tag{2.4}$$

For vectors $\mathbf{i} = (i_1,\ldots,i_r)$ in the index set, let

$$\Lambda_{r,m} = \{(i_1,\ldots,i_r) : 1 \le i_1 \le \cdots \le i_r \le m, i_1 + \cdots + i_r = m\}.$$

Then

$$B_{m,n}(P) = \sum_{r=1}^{m} \sum_{\mathbf{i}\in\Lambda_{r,m}} \omega_{m,n}(\mathbf{i}) \binom{n}{r}^{-1} \sum_{1\le j_1<\cdots<j_r\le m} h_{\mathbf{i}}(X_{j_1},\ldots,X_{j_r}) \tag{2.5}$$

where

$$\omega_{m,n}(\mathbf{i}) = \binom{n}{r}\binom{m}{i_1,\ldots,i_r}\Big/ n^m.$$

Let

$$\theta_{m,n}(P) = E_P B_{m,n}(P) = \sum_{r=1}^{m} \sum_{\mathbf{i}\in\Lambda_{r,m}} \omega_{m,n}(\mathbf{i}) E_P h_{\mathbf{i}}(X_1,\ldots,X_r). \tag{2.6}$$

Finally, let

$$\delta_m(r/m) \equiv \max\left\{|E_P h_{\mathbf{i}}(X_1,\ldots,X_r) - \theta_m(P)| : \mathbf{i} \in \Lambda_{r,m}\right\} \tag{2.7}$$

and define $\delta_m(x)$ by extrapolation on $[0,1]$. Note that $\delta_m(1) = 0$.

THEOREM 2.2. *Under the conditions of Theorem 2.1*

$$B_{m,n}(P) = \theta_{m,n}(P) + O_p(m/n)^{1/2}. \tag{2.8}$$

If further

$$\delta_m(1 - xm^{-1/2}) \to 0 \tag{2.9}$$

uniformly for $0 \le x \le M$ for all $M < \infty$ and $m = o(n)$, then

$$\theta_{m,n}(P) = \theta_m(P) + o(1). \tag{2.10}$$

Finally if

$$T_m(\mathbf{X}_1^{(i_n)}, \dots, \mathbf{X}_r^{(i_r)}, P) = T_m(\mathbf{X}_1^{(i_1)}, \dots, \mathbf{X}_r^{(i_r)}, P_n) + o_p(1) \tag{2.11}$$

whenever $\mathbf{i} \in \Lambda_{r,m}$*,* $m \to \infty$ *and* $\max\{i_1, \dots, i_r\} = O(m^{1/2})$*, then*

$$B_{m,n} = \theta_m(P) + o_p(1) \tag{2.12}$$

if $m \to \infty$ *and* $m = o(n)$*.*

The m out of n WOR bootstrap works in absolute generality while the WR bootstrap works in all examples of bootstrap failure we have addressed so far, confidence bounds for extrema, setting critical values for tests, as well as others such as estimating the kurtosis of the median – see Bickel and Sakov (2002) for instance. An extensive discussion of extensions of the WOR bootstrap is in Politis *et al.* (2002).

The advantage of the WOR bootstrap is its complete generality. Its disadvantage is that it does not connect smoothly with the ordinary n out of n bootstrap. It is intuitively clear and shown explicitly in various cases in Bickel *et al.* (1997) that when the Efron's bootstrap works, it is preferable in terms of higher order effects (beyond consistency) to the m out of n bootstraps. The WR bootstrap permits smooth extrapolation from m/n small to $m/n \simeq 1$.

Both WR and WOR methods also face·two further issues.

(i) To meaningfully apply them, the scale of $T_n(P_n, P)$ must be known or estimable, as applying the m out of n bootstrap to $n^{1/2}\{\max(X_1, \dots, X_n) - F^{-1}(1)\}$ rather than $n\{\max(X_1, \dots, X_n) - F^{-1}(1)\}$ under the assumptions of Example F2, would give us nothing since $n^{1/2}\{\max(X_1, \dots, X_n) - F^{-1}(1)\} \xrightarrow{P} 0$. Bertail *et al.* (1999) showed how this problem may be tackled using the m out of n bootstrap distributions of $T_n(P_n, P)$ for $m = 1, \dots, n$.

(ii) Knowing that $m \to \infty$, $m/n \to 0$ works tells us nothing about the magnitude of m for a particular n. Ideally, we would like a data determined choice $\widehat{m}_n$ such that the distribution $\mathcal{L}^*(T_{\widehat{m}_n}(X_1^*, \dots, X_{\widehat{m}_n}^*, P_n))$ is as close to the $\mathcal{L}_P(T_n(P_n, P))$ as possible within this family of possible estimates $\{\mathcal{L}^*(T_m(X_1^*, \dots, X_m^*, P_n))\}$. A rule having oracle properties of this type is discussed in the next section and in detail in Bickel and Sakov (2003).

3. SELECTION RULE FOR m IN THE m OUT OF n BOOTSTRAP

We develop a rule for selecting m discussed fully in Götze and Rauckauskas (2002) and Bickel and Sakov (2003). For a given $T_n(P_n, P)$, let L_n be the true distribution of T_n, $L_n \equiv \mathcal{L}_P T_n(P_n, P)$ and $L^*_{m,n}$ the corresponding bootstrap distribution $L^*_{m,n} \equiv \mathcal{L}^* T_m(P^*_m, P_n)$. Here is a description and motivation of the rule from Bickel and Sakov (2003).

Consider a sequence of m's of the form

$$m_j = \left[q^j n\right], \text{ for } j = 0, 1, 2, \ldots, \ 0 < q < 1, \tag{3.1}$$

where $[\alpha]$ denotes the smallest integer $\geq \alpha$. Here is our rule:

1. For each m_j, find $L^*_{m_j,n}$. In practice this is done by Monte-Carlo.
2. Let $\widehat{m}_n = \operatorname{argmin}_{m_j} \| L^*_{m_j,n}(\cdot) - L^*_{m_{j+1},n}(\cdot) \|_\infty$. If the difference is minimized for a few values of m_j then pick the largest among them. Denote the j corresponding to $\widehat{m}_n$ by $\widehat{\jmath}$.
3. The estimator of L is $\widehat{L} = L^*_{\widehat{m}_n,n}$.
4. Estimate θ by $\widehat{\theta}_n = \nu(\widehat{L})$ or use the quantiles of $\widehat{L}$ to construct confidence interval for θ.

Here $\|g\|_\infty = \sup_x |g(x)|$, the sup distance (Kolmogorov–Smirnov) for instance, though it can be some other suitable measure of deviation. For instance, in testing, a topic we do not pursue in this paper, comparison using p-values is more appropriate (see Bickel and Sakov, 2002).

Here is the rationale behind this rule. In essentially all examples the failure of the n bootstrap is of the following type: $L^*_{n,n}$, viewed as a probability distribution on the space of all probability distributions, does not converge to a point mass at the correct limit L but rather converges to a nondegenerate distribution, call it $\mathcal{L}_1$, on that space. If $m \to \infty$, $m/n \to \lambda$, $0 < \lambda < 1$, one gets convergence to a nondegenerate distribution, $\mathcal{L}_\lambda$, which is typically different from $\mathcal{L}_1$. We expect $\mathcal{L}_0 = L$. On the other hand, if m is fixed, $L^*_{m\,n}$ typically converges to a degenerate distribution concentrated at $L_m(F) = \mathcal{L}_F(T_m(X_1, \ldots, X_m, F))$. The motivation of the rule should now be clear if $m \mapsto L_m(F)$ and $\lambda \mapsto \mathcal{L}_\lambda$ are one-to-one. We reconsider Example F3 where $T_n(X_1, \ldots, X_n) = \sqrt{n}\bar{X}_n$, and where $X_1, \ldots, X_n$ are *iid* with zero mean and variance $\sigma^2(F) < \infty$. We use the subscript on $\bar{X}$ below to indicate sample size. Then, $\sqrt{m}(\bar{X}^*_m - \bar{X}_n) \overset{\mathcal{L}}{\Rightarrow} N(0, \sigma^2(F))$ with probability one as

$m \to \infty$. But, this implies that $\sqrt{m}\bar{X}_m^*$ behaves like $N(\sqrt{m}\bar{X}_n, \sigma^2(F))$. Writing $\sqrt{m}\bar{X}_n = \sqrt{m/n}\sqrt{n}\bar{X}_n$, we see that if $m/n \to \lambda > 0$, $\sqrt{m}\bar{X}_n \overset{\mathcal{L}}{\Rightarrow} N(0, \lambda\sigma^2(F))$, *i.e.*, that $\mathcal{L}_\lambda$ is the random distribution given by $N(\sqrt{\lambda}Z, \sigma^2(F))$ where $Z \sim N(0, \sigma^2(F))$. Note that $\mathcal{L}_\lambda$ is degenerate and equal to $\mathcal{L}_1 = N(0, \sigma^2(F))$ if and only if $\lambda = 0$. Furthermore, $\lambda \mapsto \mathcal{L}_\lambda$ is one-to-one. Moreover, by a theorem of Kagan *et al.* (1973), in this set-up $\sqrt{m_1}\bar{X}_{m_1} \overset{d}{=} \sqrt{m_2}\bar{X}_{m_2}$ for $m_1 \neq m_2$ if and only if X_1 is Gaussian. Thus, our rule should work if F is not Gaussian. In fact, it should and does work even if X_1 is Gaussian since then any choice of m will work.

It is shown in Bickel and Sakov (2003) that the $\widehat{m}_n$ selected by this rule indeed has $\widehat{m}_n \xrightarrow{P} \infty$, $\widehat{m}_n/n \xrightarrow{P} 0$ in typical situations where the Efron's bootstrap is inconsistent. More importantly, if we let m_n be the "oracle choice", the value of m minimizing the distance, $d(L^*_{m,n}, L_n)$ between $L^*_{m,n}$ and L_n the distribution we are estimating, then $d(L^*_{\widehat{m}_n,n}, L_n)$ and $d(L^*_{m_n,n}, L_n)$ are of the same order. Essentially, $\widehat{m}_n$ is a correct choice to second order. These results, their application to confidence bounds for high quantiles, for instance, $F^{-1}(1 - 1/n)$ for F the distribution function under P, and simulations are in Bickel and Sakov (2003).

Similar results for applications such as estimating test quantiles may be found in Götze and Rauckauskas (2002). Moreover, Bickel and Sakov (2003) also demonstrate that if indeed $m_n/n \xrightarrow{P} 1$, that is the Efron's bootstrap is actually best, $\widehat{m}_n/n \xrightarrow{P} 1$ as well. So this selection rule is potentially successful across the whole possible scale of misbehaviour and optimal behaviour of the Efron's bootstrap.

There is a class of situations in which the m out of n bootstrap works, but even at its best is suboptimal. A possible fix is extrapolation, a technique discussed in Bickel and Sakov (2002). We do not dwell on this here but instead in the next section discuss a second unorthodox variant of the Efron's bootstrap which "automatically" corrects the inconsistency involved in the naive use of the n out of n bootstrap.

4. The "Confidence Band" Bootstrap

As we noted earlier, using the ordinary Efron's bootstrap to set critical values of $\bar{X}$ in testing $H : E_P X = 0$ *vs.* $K : E_P X > 0$ gives incorrect results. In this case, the malfunction is expected. In general, when testing $H : P \in \mathcal{P}_0$ with a test statistic $T_n(P_n)$, the more appropriate thing to do is to find an estimate $\widehat{P}_0$ of $P \in \mathcal{P}_0$ which is consistent for $P_0 \in \mathcal{P}_0$ *i.e.* $\mathcal{L}^*_{\widehat{P}_0}(T_n(P_n^*)) \to \mathcal{L}_{P_0}$ in P_0 probability if $\mathcal{L}_{P_0}(T_n(P_n)) \to \mathcal{L}_{P_0}$. Here, $\mathcal{L}^*_{\widehat{P}_0}$ refers to sampling $X_1^*, \dots, X_n^*$ *iid* from $\widehat{P}_0$. For instance, in the example we have discussed, a natural choice of $\widehat{P}_0$ is the empirical distribution of $X_1 - \bar{X}, \ldots, X_n - \bar{X}$ which belongs to $\mathcal{P}_0$ since it has expectation 0. This way of proceeding is the natural generalization of what is done for parametric $\mathcal{P}_0 = \{P_\theta : \theta \in \Theta_0 \subset R^d\}$: find $\widehat{\theta}_0$, the MLE under $\mathcal{P}_0$, and simulate $P_{\widehat{\theta}_0}$.

However, for semiparametric (and even parametric) $\mathcal{P}_0$ finding an appropriate $\widehat{P}_0$ and simulating from it may be difficult. Here is an example from Bickel and Ren (1996, 2002).

EXAMPLE C1 (*Double censoring*). Let $X_1, \ldots, X_n$ be *iid* with common distribution that of X where $X = (Y, \delta)$ and Y and δ are defined as follows. Let Z have a distribution F and be independent of (L, U) where $L < U$. Let (L, U) have a joint distribution G. Then

$$Y = \begin{cases} Z, \; L \le Z \le U, \\ L, \; Z < L, \\ U, \; Z > U \end{cases} ; \quad \delta = \begin{cases} 0, \; L \le Z \le U, \\ -1, \; Z < L, \\ 1, \; Z > U. \end{cases}$$

If we parametrize the members of the model as $P_{(F,G)}$, the hypothesis of interest is $H : F = F_0$. The parameter of interest defining $H : F = F_0$ is $T(P_{(F,G)}) = \int (F - F_0)^2 dF_0$. The test statistic is $T_n(P_n) = n \int (\widehat{F}_n - F_0)^2 dF_0$ where $\widehat{F}_n$ is the NPMLE of F, which can be thought of as $\tau(\sqrt{n}(\widehat{F}_n - F_0))$. Algorithms for finding $\widehat{F}_n$ are given by Turnbull (1974).

The problem in finding $\widehat{P}_0 \equiv P_{(F_0, \widehat{G})}$ here is that the censoring mechanism is quite unknown and L and U are never observed together so that we are unable to estimate G by $\widehat{G}$ and thus generate observations as we would do naturally by using $\widehat{P}_0$. Even if only right censoring is at issue, $\widehat{P}_0$, obtained by computing the Kaplan-Meier estimate $\widehat{G}$ of G and then obtaining observations from $P_{(F_0, \widehat{G})}$ is not necessarily simple if F_0 is not a standard distribution, and rejective sampling or a similar method need to be used.

The m out of n bootstrap is, as we noted, an alternative but, as is shown in Bickel and Ren (2002), power loss over the alternative we describe below is an unavoidable consequence. The alternative, introduced implicitly by Beran (1986) and reproposed with many examples in Bickel and Ren (2002) is to consider a confidence band problem. In this problem and quite generally as Bickel and Ren (2002) show, the semiparametric hypothesis testing problem we consider is of the form $H : \; T(P) = 0$ where T can be a function. Thus, in Example C1, let $F(P)$ be the functional corresponding to the NPMLE of F, due to Turnbull (1974) in this case. Then, $T(P) = F(P) - F_0$. Under regularity conditions $\sqrt{n}(F(P_n) - F(P))$ has a limiting Gaussian "distribution" whatever be P. Thus, to set critical values for a statistic, $T_n(P_n)$ of the form $\tau(\sqrt{n}(F(P_n) - F_0))$ such that $\tau(0) = 0$, as we have considered in Example C1, it is enough to obtain an estimate of the distribution of $\sqrt{n}(F(P_n) - F(P))$ to sample from. The natural procedure is to use the Efron's bootstrap distribution not of $\tau(\sqrt{n}(F(P_n) - F_0))$ but rather that of $\tau(\sqrt{n}(F(P_n^*) - F(P_n)))$. This gives us the right answer not only for $P \in \mathcal{P}_0$ where the limit of the bootstrap distributions is the same as the limiting distribution of $\tau(\sqrt{n}(F(P_n) - F_0)$ but also when $F(P) \neq F_0$ when $\tau(\sqrt{n}(F(P_n) - F_0))$ tends to ∞. So in general, our principle is that if we can frame H as $H : T_n(P) = 0$ and we consider a statistic $\tau(T_n(P_n))$, then we should obtain a critical value by using the ordinary Efron's (n out of n) bootstrap distribution, $\mathcal{L}^*(\tau(T_n(P_n^*) - T_n(P_n)))$. This idea is developed in a number of examples in Bickel and Ren (2002).

In conclusion, when it can be used, the confidence band bootstrap is a simpler and theoretically better approach than that of the m out of n bootstrap. However, the latter as is discussed in Bickel and Ren (2002) and Bickel and Sakov (2003) is applicable much more broadly and when coupled with the selection rule given in Section 3, will have oracular properties (within the class of m out of n bootstraps).

REFERENCES

ATHREYA, K. B. AND FUKUCHI, J. (1994). "Bootstrapping extremes of iid random variables", *Proceedings of Conference on Extreme Value Theory (NIST).*

BARLOW, R. E., BARTHOLOMEW, D. J., BRENNER, J. M. AND BRUNK, H. D. (1972). *Statistical Inference under Order Restrictions*, Wiley, London.

BERAN, R. (1986). "Simulated power functions", *The Annals of Statistics*, **14**, 151–173.

BERTAIL, P., POLITIS, D. N. AND ROMANO, J. P. (1999). "Subsampling estimators with unknown rate of convergence", *Journal of the American Statistical Association*, **94**, 569–579.

BICKEL, P. J. AND DOKSUM, K. (2004). *Mathematical Statistics*, Vol. 2 (to appear), Prentice Hall.

BICKEL, P. J. AND FREEDMAN, D. A. (1981). "Some asymptotic theory for the bootstrap", *The Annals of Statistics*, **9**, 1196–1217.

BICKEL, P. J., GÖTZE, F. AND VAN ZWET, W. R. (1997). "Resampling fewer than n observations : Gains, losses, and remedies for losses", *Statistica Sinica*, **7**, 1–31.

BICKEL, P. J. AND REN, J. J. (1996). "The m out of n bootstrap and goodness of fit tests with doubly censored data", In *Robust Statistics, Data Analysis and Computer Intensive Methods. Lecture Notes in Statistics.* (H. Rieder, ed.), Springer-Verlag, New York.

BICKEL, P. J. AND REN, J. J. (2002). "The bootstrap in hypothesis testing", *IMS Lecture Notes Monograph Series*, Vol. 36, Institute of Mathematical Statistics, Beachwood.

BICKEL, P. J. AND SAKOV, A. (2002). "Extrapolation and the bootstrap", *Sankhyā*, **64**, 640–652.

BICKEL, P. J. AND SAKOV, A. (2003). "On the choice of m in the m out of n bootstrap estimation problems", *The Annals of Statistics*, submitted.

EFRON, B. AND TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall, London, New York.

FREEDMAN, D. A. (1981). "Bootstrapping regression models", *The Annals of Mathematical Statistics*, **12**, 1218–1228.

GÖTZE, F. (1993). *Bulletin IMS.*

GÖTZE, F. AND RAUCKAUSKAS, A. (2002). "Adaptive choice of bootstrap sample sizes : State of the art in probability and statistics", *IMS Lecture Notes Monograph Series*, Vol. 36, Institute of Mathematical Statistics, Beachwood.

HALL, P. (1992). *The Bootstrap and Edgeworth Expansion*, Springer–Verlag, New York.

KAGAN, A. M., LINNIK, Y. V. AND RAO, C. R. (1973). *Mathematical Statistics*, Wiley, New York.

MAMMEN, E. (1992). *When Does Bootstrap Work?*, Springer–Verlag, New York.

POLITIS, D. N. AND ROMANO, J. P. (1994). "A general theory for large sample confidence regions based on subsamples under minimal assumptions", *The Annals of Mathematical Statistics*, **22**, 2031–2050.

POLITIS, D. N., ROMANO, J. P. AND WOLF, M. (2002). *Subsampling*, Springer, New York.

TURNBULL, B. W. (1974). "Nonparametric estimation of a survivorship function with doubly censored data", *Journal of the American Statistical Association*, **69**, 169–173.

VAN DER VAART, A. AND WELLNER, J. (1996). *Weak Convergence and Empirical Processes*, Springer, New York.