THE BOOTSTRAP IN HYPOTHESIS TESTING

PETER J. BICKEL[1] AND JIAN-JIAN REN[2]

University of California at Berkeley and Tulane University

We propose a unifying principle which identifies a very broad class of hypotheses and statistics for which a suitable application of the n out of n bootstrap yields asymptotically correct critical values and power for contiguous alternatives. We also show that this attractive principle can fail in situations which the m out of n bootstrap can deal with (Bickel, Götze and van Zwet, 1997)(BGvZ). We formalize the m out of n bootstrap theory for testing and show that under mild conditions, it provides correct significance level, asymptotic power under contiguous alternatives, and consistency. We conclude with simulation results supporting the asymptotics.

AMS subject classifications: 62G09, 62G10.
Keywords and phrases: Censored data, change points, empirical processes, goodness of fit, Hadamard differentiability, manifolds, U statistics.

1 Introduction

It is logically clear but not always noted that the usual nonparametric bootstrap (the n out of n bootstrap) fails in setting critical values for test statistics in hypothesis testing. The problem is that hypothesis restrictions are not reflected adequately by the empirical distribution when one is resampling as many observations as one has in the sample. For example, Freedman (1981) points out that in setting confidence intervals for the usual slope estimate for regression through the origin, one must resample not the residuals but the residuals centered at their mean. If one considers setting confidence bands as the dual of hypothesis testing, a moment's thought will show that not centering the residuals is tantamount to not imposing the model requirement for the hypothesis tests that the expectation of the error is 0. For more recent examples, see Härdle and Mammen (1993), Mammen (1992) and Bickel, Götze and van Zwet (1997) (BGvZ). BGvZ note that the m out of n bootstrap, $m \to \infty$, $\frac{m}{n} \to 0$, is in principle usable. In particular, Bickel and Ren (1996) study the following situation: testing for goodness of fit with doubly censored data where the usual bootstrap as usual fails and finding a distribution approximating the truth under H_0 is difficult. They propose using the m out of n bootstrap to set the critical value of the test

[1]Research partially supported by NSF Grant DMS 9504955.
[2]Research partially supported by NSF Grants DMS 9510376 and 9626535/9796229.

and show that the proposed testing procedure is asymptotically consistent and has correct power against $\sqrt{n}$-alternatives.

There have been a number of papers in the literature detailing modifications of the bootstrap for correct use in testing, see Beran (1986), Beran and Millar (1987), Hinkley (1987, 1988), Romano (1988, 1989), among others. In particular, Hinkley indicated quite generally that bootstrapping from a distribution obeying the constraints of the hypothesis which is closest in some metric to the empirical distribution should give asymptotically correct critical values. Unfortunately, this requires an exercise in ingenuity in most cases, and as has been frequently noted, say, Shao and Tu (1995) for example, that it may in practice be very difficult to construct such a distribution. Romano showed that in an interesting class of situations, including testing goodness of fit to parametric composite hypothesis and independence, there was a natural definition of a distribution in the null hypothesis H_0 closest to the empirical, and that, for natural test statistics, bootstrapping from this distribution would yield asymptotically appropriate critical values. In a prescient paper, Beran (1986) gave two general principles for construction of tests of abstract hypotheses in the presence of abstract nuisance parameters and estimation of the power functions of such tests.

In Section 2, we propose a unifying principle which identifies a very broad class of hypotheses and statistics including all those considered by Romano (1988) for which a suitable application of the n out of n bootstrap yields asymptotically correct critical values, power for contiguous alternatives, and consistency under mild conditions. We state a general theorem and apply it in eight examples including all those of Romano, those of Bickel and Ren (1996), a test for change-point (Matthews, Farewell and Pyke, 1985) with censored data, and a number of others. This result, Theorem 2.1, applies only to test statistics which are regular in the sense of stabilizing on the $n^{-1/2}$ scale under the hypothesis. We then in Theorem 2.2 extend Theorem 2.1 to a broader class of statistics based on estimates of irregular parameters such as densities. Moreover, we show that our proposed unifying principle can fail in situations which the m out of n bootstrap can deal with.

Our unifying principle, though not our point of view, can be viewed as a particular case of one of Beran's two approaches, even as Hinkley's work corresponds to the other. However, that part of Beran's formulation which is relevant to the principle we state emphasized construction of tests from confidence region for abstract parameters in the presence of nuisance parameters rather than the setting of critical values for natural test statistics. Perhaps for this reason, the abstract point of view which obscured the rather simple geometrically based special case we focus on and the general conditions whose checking is usually the heart of the matter, the broad applicability of his argument was not appreciated (even by us until a referee

brought his paper to our attention). We focus here on checkable conditions and examples.

In Section 3, we state and prove a theorem showing that the m out of n bootstrap is an approach that generally provides correct significance level, asymptotic power under contiguous alternatives, and consistency. This is essentially a formalization of the discussion of BGvZ.

We close with simulations and a brief appendix indicating where the regularity conditions for the examples can be found.

2 A general approach to defining semiparametric hypotheses

For simplicity, we start this section with the case where the data $X_1, \ldots, X_n$ are independently and identically distributed (i.i.d.) taking values in $\mathcal{X}$, usually R^k, with an unknown distribution function (d.f.) $F \in \mathcal{F}$. However, it should be apparent from our discussion that our approach is more generally applicable.

Suppose that we want to test

$$H_0 : F \in \mathcal{F}_0 \text{ vs. } H_1 : F \notin \mathcal{F}_0. \tag{2.1}$$

We begin with considering the case that X takes on $k+1$ values $x_0, x_1, \ldots, x_k$ only. Thus $\mathcal{F}$ is parametrized by

$$p_j = P\{X = x_j\},\ 0 \le j \le k,$$

$$\boldsymbol{p} \in I\!P = \left\{ \boldsymbol{p} \in R^k;\ p_j > 0,\ 1 \le j \le k,\ \sum_{j=1}^{k} p_j < 1 \right\}.$$

A hypothesis $\mathcal{F}_0$ is then described by, say, $\{\boldsymbol{h}(\boldsymbol{\theta}) \in I\!P;\ \boldsymbol{\theta} \in R^q,\ q < k\}$, where $\boldsymbol{\theta} \to \boldsymbol{h}(\boldsymbol{\theta})$ is $1-1$. If $\boldsymbol{h}(\boldsymbol{\theta})$ is continuously differentiable and $(\partial h_i(\boldsymbol{\theta})/\partial \theta_j)_{k \times q}$ is of full rank, then $\boldsymbol{h}$ is an *embedding* of R^q in R^k (Vaisman, 1984, page 11, 13 and 15). This means that for any $\boldsymbol{p}_0 \in \mathcal{F}_0 = \boldsymbol{h}(R^q) \cap I\!P$, there exist open sets U_{p_0} and U_0 in R^k and a differentiable function $\boldsymbol{\eta}_0 : U_0 \to R^{k-q}$, such that $\boldsymbol{p}_0 \in U_{p_0}$ and $U_{p_0} \cap \mathcal{F}_0 = \{\boldsymbol{p} \in U_0 \mid \boldsymbol{\eta}_0(\boldsymbol{p}) = \boldsymbol{0}\}$. The map ("atlas") $\boldsymbol{\eta}_0$ can in many cases of interest be pieced together consistently to a single $\boldsymbol{\eta}_0$ such that

$$\mathcal{F}_0 = \{\boldsymbol{p} \in I\!P \mid \boldsymbol{\eta}_0(\boldsymbol{p}) = \boldsymbol{0}\}.^3 \tag{2.2}$$

Thus, if the random sample $X_1, \ldots, X_n$ is from some $\boldsymbol{p} \in I\!P$, which is in a neighborhood U_{p_0} of some $\boldsymbol{p}_0 \in \mathcal{F}_0$, and if $\hat{\boldsymbol{\eta}} \equiv \boldsymbol{\eta}_0(\hat{\boldsymbol{p}})$, where $\hat{\boldsymbol{p}} = (N_1/n, N_2/n, \ldots, N_k/n)^T$ is the empirical distribution (vector of frequencies) of $X_1, \ldots, X_n$, typical tests are of the form (or asymptotically equivalent to tests of the form): Reject if $\tau(\sqrt{n}\,\hat{\boldsymbol{\eta}})$ is large, where the function

[3] Even if the "atlas" can not be reduced to a single function, we still can naturally base a test on $\boldsymbol{\eta}_0(\hat{\boldsymbol{p}})$ for $\hat{\boldsymbol{p}}$ as above and $\hat{\boldsymbol{p}}_0$ as the member of $\mathcal{F}_0$ closest to $\hat{\boldsymbol{p}}$.

$\tau : R^{k-q} \to R^+$ is continuous with $\tau(\boldsymbol{t}) = 0$, iff $\boldsymbol{t} = \mathbf{0}$. Typically, τ is equivalent to a norm on R^{k-q}.

For instance, the usual Wald test is to use $n\hat{\boldsymbol{\eta}}^T\hat{\Sigma}^{-1}\hat{\boldsymbol{\eta}}$, where $\hat{\Sigma}$ is an estimate of the covariance matrix $\Sigma(\boldsymbol{p})$ of $\hat{\boldsymbol{\eta}}$. This is equivalent to using $\tau(\boldsymbol{x}) = \boldsymbol{x}^T\Sigma_0^{-1}\boldsymbol{x}$. In this situation, we can bootstrap parametrically in one of two ways:

(a) Estimate $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}_0$, the maximum likelihood estimator (MLE) under $H_0 : \boldsymbol{\eta}_0 = \mathbf{0}$, then use the appropriate percentile of the distribution of $\tau(\sqrt{n}\,\hat{\boldsymbol{\eta}}^*)$ as the critical value, where $X_1^*, \ldots, X_n^*$ are i.i.d. $\boldsymbol{p}(\hat{\boldsymbol{\theta}}_0)$;

(b) Note that $\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0)$ and $\sqrt{n}\,\hat{\boldsymbol{\eta}}$ have the same distributions under H_0 and use the appropriate percentile of the distribution of $\tau(\sqrt{n}(\hat{\boldsymbol{\eta}}^* - \hat{\boldsymbol{\eta}}))$ as the critical value, but where now $X_1^*, \ldots, X_n^*$ may be obtained from the 'nonparametric' bootstrap, i.e., i.i.d. $\hat{\boldsymbol{p}}$.

If $\hat{\boldsymbol{\theta}}_0$ is uniformly consistent on Θ, it follows from, for example, a theorem of Rao (1973, page 360–362) that these bootstraps are both valid. (Note that $\boldsymbol{p}(\hat{\boldsymbol{\theta}}_0)$ can be used instead of $\hat{\boldsymbol{p}}$ in case (b).)

If X does not have finite support, the corresponding conditions for characterization of an embedding in Hilbert space are more involved. Nonetheless, as we shall see by example below, the equivalence (2.2) holds quite broadly.

Sufficient conditions for use of bootstrap (b) are easily given. Suppose that for hypothesis (2.1), there exists $T : \mathcal{F} \to \mathcal{T}$, where $\mathcal{T}$ is a Banach space, possibly R^p but often a function space such as $D[R^p]$, such that

$$\mathcal{F}_0 = \{F;\, T(F) = 0\}. \tag{2.3}$$

It is often convenient to think of both $\mathcal{F}$ and $\mathcal{T}$ as subsets of spaces of finite signed measures defined on spaces of bounded functions $\mathcal{H}_x, \mathcal{H}_y$ on $\mathcal{X}$ and another space $\mathcal{Y}$ and identify F as a member of $l_\infty(\mathcal{H}_x), T(F) = G$ as a member of $l_\infty(\mathcal{H}_y)$ via,

$$F(h) = \int h(x)\, dF(x),\ G(r) = \int r(y)\, dG(y). \tag{2.4}$$

We shall throughout assume that measurability technicalities are dealt with by the Hoffman-Jørgensen approach — see van der Vaart and Wellner (1996).

Let $\hat{F}_n$ denote the empirical distribution of $X_1, \ldots, X_n$ and $\tau : \mathcal{T} \to R^+$ be continuous with $\tau(\boldsymbol{t}) = 0$ iff $\boldsymbol{t} = \mathbf{0}$. Tests for (2.1) are naturally based on rejecting H_0 for large $\tau(\sqrt{n}T(\hat{F}_n))$ (provided that $\sqrt{n}(T(\hat{F}_n) - T(F))$ is well behaved). In analogy to the multinomial situation, it seems natural to use either

(a) The quantiles of the distribution of $\tau(\sqrt{n}T(\hat{F}_n^*))$, where $X_1^*, \ldots, X_n^*$ are i.i.d. from $\hat{F}_0 \in \mathcal{F}_0$, which is a 'uniformly consistent' estimate of F under H_0; or

(b) The quantiles of the nonparametric n out of n bootstrap distribution of $\tau(\sqrt{n}(T(\hat{F}_n^*) - T(\hat{F}_n)))$, where $\hat{F}_n^*$ is the empirical distribution of $X_1^*, \ldots, X_n^*$, a sample from $\hat{F}_n$;

as critical values for $\tau(\sqrt{n}T(\hat{F}_n))$. In the framework of Beran (1986), this can be viewed as using the test: Accept iff $0 \in \mathbb{C}(\hat{F}_n)$, where $\mathbb{C}(\hat{F}_n)$ is the asymptotic $1-\alpha$ confidence region $\{t; \tau(\sqrt{n}(T(\hat{F}_n) - t)) \leq d_n(\hat{F}_n)\}$ and $d_n(F)$ is the $1-\alpha$ quantile of the distribution of $\tau(\sqrt{n}(T(\hat{F}_n) - T(F)))$. We shall give sufficient conditions for the validity of alternative (b) in this abstract framework below, but before doing so we give some examples where (2.3) applies.

Example 1 *Goodness of fit to a single hypothesis.* Here $\mathcal{F}_0 = \{F_0\}, \mathcal{F}$ is all distributions and we can clearly take $\mathcal{T} = l_\infty(\mathcal{X})$, finite signed measures on $\mathcal{X}$ and $T(F) = F - F_0$. Possible τ's are $\tau(\mu) = \|\mu\|_\infty$, where $\|\cdot\|_\infty = \sup\{|\mu(h)| : h \in \mathcal{H}_x\}$ for $\mathcal{X}$ suitable $\mathcal{H}_x$. For example, $\mathcal{X} = R$, $\mathcal{H}_x = \{1_{(-\infty,t)}; t \in R\}$ gives the Kolmogorov–Smirnov test. Another possibility is weighted averages of $\mu^2(h)$ over $\mathcal{H}_x$. Thus, $\tau(\mu) = \int(\mu(-\infty, x))^2 dF_0(x)$ leads to the Cramér-von Mises test. Option (b) corresponds to using the bootstrap distribution of $\tau(\sqrt{n}(\hat{F}_n^* - \hat{F}_n))$, while (a) leads to simulating from F_0. □

Example 2 *Goodness of fit to a composite hypothesis.* Here $\mathcal{F}_0 = \{F_\theta; \theta \in \Theta\}$, $\theta \in R^d$, say, and $\mathcal{F}_0$ is a regular parametric model. Suppose that $\theta(\hat{F}_n) \in \Theta$ is a regular estimate of θ in the sense of Bickel, Klaassen, Ritov and Wellner (1993) (BKRW) where $\theta : \mathcal{F} \to \Theta$ is a parameter. For instance $\theta(F) = \arg\min \|F - F_\theta\|_\infty$ may be a possibility. Again we can take $\mathcal{F} \subset l_\infty(\mathcal{H}_x)$ and

$$T(F) = F - F_{\theta(F)}.$$

Note that we could take $\theta(F)$ as any parameter defined on $\mathcal{F}$ such that $\theta(F_\theta) = \theta$.

This example figures prominently in Romano (1988). There he considered $\mathcal{F}_0$ describable by $\mathcal{F}_0 = \{F; F = \gamma(F)\}$ and recommended scheme (a), resampling from $\gamma(\hat{F}_n)$ for statistic $\|\hat{F}_n - \gamma(\hat{F}_n)\|$ and $\gamma(F) = F_{\theta(F)}$. Our scheme simply rewrites $F = \gamma(F)$ as $F - \gamma(F) = 0$. However we prescribe bootstrapping from the empirical for statistic $\sqrt{n}\|\hat{F}_n - \gamma(\hat{F}_n) - F + \gamma(F)\|$. That is we use the bootstrap distribution of $\sqrt{n}\|\hat{F}_n^* - \gamma(\hat{F}_n^*) - \hat{F}_n + \gamma(\hat{F}_n)\|$. □

Example 3 *Tests of location.* Suppose $\mathcal{X} = R^q$ and T is a location parameter

$$T(F(\cdot - \theta_0)) = T(F) + \theta_0 \tag{2.5}$$

for all $\theta_0 \in R^q$. Let $\mathcal{F}_0 = \{F;\ T(F) = 0\}$. Thus if $T(F) = \int x\, dF$, this is the hypothesis that the population mean of F is 0. If $T(F) = F^{-1}\left(\frac{1}{2}\right)$, then this is the hypothesis that the population median is 0. We can similarly consider trimmed means etc. In fact our discussion applies to scale parameters and more generally transformation parameters — see BKRW (1993) — but we do not pursue this. In this case our prescription is to use the bootstrap distribution of $\sqrt{n}(T(\hat{F}_n^*) - T(\hat{F}_n))$. Here prescriptions (a) and (b) coincide since the distribution of $\sqrt{n}(T(\hat{F}_n^*) - T(\hat{F}_n))$ under H_0 is by (2.5) the same as that of $\sqrt{n}\,T(\tilde{F}_n^*)$ where $\tilde{F}_n^*$ is the empirical distribution of a sample from $\hat{F}_n(\cdot - T(\hat{F}_n))$. Equivalently say in the case of the mean the bootstrap distribution of $\bar{X}_n^* - \bar{X}_n$ is the same as the distribution of $\tilde{\bar{X}}_n$, the mean of a resample from the residuals $X_1 - \bar{X}, \ldots, X_n - \bar{X}$. The latter (a) form is the prescription of Freedman (1981) and Romano (1988), and the special case of the mean is Example 2 of Beran (1986). □

We now turn to some simple results.

Suppose $\mathcal{T}$ is a subset of a space of finite signed measures with $\bar{\mathcal{T}}$ viewed as a subset of the Banach space $l_\infty(\mathcal{H}_y)$ as above.

Suppose T is extendable to $\bar{\mathcal{F}}$ and,

(A1) T is Hadamard differentiable at all $F_0 \in \mathcal{F}_0$ as a map from $(\bar{\mathcal{F}}, \|\cdot\|_\infty)$ to $(\bar{\mathcal{T}}, \|\cdot\|_\infty)$ with derivative $\dot{T} : \bar{\mathcal{T}} \to l_\infty(\mathcal{H}_y)$, a continuous linear transformation, and $\bar{\mathcal{F}}$ a closed linear space containing $\mathcal{F}$. That is

$$\sup\{\|T(F_0 + \lambda\Delta) - T(F_0) - \lambda\dot{T}(F_0)\Delta\|;\ \Delta \in K\} = o(\lambda) \tag{2.6}$$

where K is any compact subset of $l_\infty(\mathcal{H}_x)$ and $\lambda \to 0$.

(A2) $\sqrt{n}(\hat{F}_n - F_0) \Rightarrow Z_{F_0}$ in the sense of weak convergence for probabilities on $l_\infty(\mathcal{H}_x)$ given by Hoffman-Jørgensen and $P\{Z_{F_0} \in \bar{\mathcal{F}}\} = 1$ for all $F_0 \in \mathcal{F}_0$.

Theorem 2.1 *Under* (A1) *and* (A2), *for all* $F_0 \in \mathcal{F}_0$,

$$\sqrt{n}(T(\hat{F}_n) - T(F_0)) \Rightarrow \dot{T}(F_0)Z_{F_0} \tag{2.7}$$

and with probability 1,

$$\sqrt{n}(T(\hat{F}_n^*) - T(\hat{F}_n)) \Rightarrow \dot{T}(F_0)Z_{F_0}. \tag{2.8}$$

Proof By Giné and Zinn (1990), (A2) implies that

$$\sqrt{n}(\hat{F}_n^* - \hat{F}_n) \Rightarrow Z_{F_0} \tag{2.9}$$

with probability 1. Now apply a standard argument. By Hadamard differentiability

$$\sqrt{n}(T(\hat{F}_n) - T(\hat{F}_0)) = \dot{T}(F_0)\sqrt{n}(\hat{F}_n - F_0) + o_p(1) \tag{2.10}$$

$$\sqrt{n}(T(\hat{F}_n^*) - T(F_0)) = \dot{T}(F_0)\sqrt{n}(\hat{F}_n^* - F_0) + o_p(1) \tag{2.11}$$

(2.10) yields (2.7) and subtracting (2.10) from (2.11) yields (2.8). ∎

Now letting $\mathcal{L}_0$ be the distribution of $\tau(\dot{T}(F_0)Z_{F_0})$, we have the following corollary.

Corollary 2.1 *Under the assumptions of Theorem 2.1, if $\mathcal{L}_0$ is continuous, and respectively, $C_\alpha^{*(n)}$ and C_α^0 are the $(1-\alpha)$-quantiles of $\tau(\sqrt{n}[T(\hat{F}_n^*) - T(\hat{F}_n)])$ and $\mathcal{L}_0$, then as $n \to \infty$,*

$$P\{\tau(\sqrt{n}T(\hat{F}_n)) \geq C_\alpha^{*(n)} \mid H_0\} \to \alpha. \tag{2.12}$$

In fact, as $n \to \infty$,

$$P\{[\tau(\sqrt{n}T(\hat{F}_n)) \geq C_\alpha^{*(n)}]\Delta[\tau(\sqrt{n}T(\hat{F}_n)) \geq C_\alpha^0] \mid H_0\} \to 0. \tag{2.13}$$

If $\{F_n\}$ is a sequence of alternatives contiguous to $F_0 \in \mathcal{F}_0$, then (2.13) continues to hold with P replaced by P_n corresponding to F_n, and hence the power functions for the tests using $C_\alpha^{*(n)}$ and $C_\alpha^{(0)}$ are the same.

If (A1) and (A2) hold for all $F \in \mathcal{F}$ not just $\mathcal{F}_0$ and $\tau(\boldsymbol{t}) \to \infty$, as $\|\boldsymbol{t}\|_\infty \to \infty$, then the test based on $C_\alpha^{*(n)}$ is consistent for all $F \notin \mathcal{F}_0$.

Proof (2.12) and (2.13) follow from $C_\alpha^{*(n)} \xrightarrow{P} C_\alpha^{(0)}$ for all $F_0 \in \mathcal{F}_0$, an immediate consequence of the theorem and Polya's theorem. Contiguity preserves convergence in probability to constants so that equivalence of the power functions follows. Finally consistency follows since under the assumption $C_\alpha^{*(n)}$ converges in probability under F to the $(1-\alpha)$-quantile of $\mathcal{L}_F(\tau(\dot{T}(F)Z_F))$. But,

$$\|\sqrt{n}T(\hat{F}_n)\|_\infty = \|\sqrt{n}(T(\hat{F}_n) - T(F)) - \sqrt{n}T(F)\|_\infty \xrightarrow{P} \infty$$

since the first term in the norm is tight while the second term has norm of the order $\sqrt{n}$ since $T(F) \neq 0$. ∎

The examples 1-3 cited above all satisfy our assumptions essentially under the mild regularity conditions needed to justify that the test statistics in question have a limit law under H_0. We discuss the conditions briefly in the appendix. Now we turn to some further examples and a mild extension.

Our next example falls outside of the Romano domain.

Example 4 *Goodness of fit test of a lifetime distribution under censoring.* Suppose that for a desired observation T_i, there are right censoring variable C_i and left censoring variable B_i such that T_i is independent from (B_i, C_i) and that the available observations are in the form $X_i = (Y_i, \delta_i)$, where in

the right censored sample case, we have $Y_i = \min\{T_i, C_i\}, \delta_i = I\{T_i \leq C_i\}$, and in the doubly censored sample case (Turnbull, 1974), we have $Y_i = \max\{\min\{T_i, C_i\}, B_i\}, \delta_i = I\{B_i < T_i \leq C_i\} + 2I\{T_i > C_i\} + 3I\{T_i \leq B_i\}$ with $P\{B_i < C_i\} = 1$. Let G be the distribution function of T_i, then in this frame work the goodness of fit test H_0: $G = G_0$, for a given G_0, is important. We write F as the distribution of $X = (Y, \delta)$. Then if G is identifiable, we have $G = \phi(F)$ with $\hat{G}_n = \phi(\hat{F}_n)$ to be the nonparametric maximum likelihood estimate (NPMLE) for G (see Bickel and Ren, 1996). Thus, we can take $T(F) = \phi(F) - G_0 = G - G_0$. Although $T(\cdot)$ is not Hadamard differentiable here, prescription (b) says to use the bootstrap distribution of $\tau(\sqrt{n}(\hat{G}_n^* - \hat{G}_n))$. As Bickel and Ren (1996) point out, it is difficult to fulfill prescription (a) in this case for doubly censored data since it is not clear what to use as the member of $\mathcal{F}_0$ from which we should resample. We will return to this example subsequently in Section 4. □

Example 5 *U statistics.* A natural generalization of Example 3 is testing $H_0 : T(F) = 0$ where $T(F) = E_F\psi(X_1, \ldots, X_k)$, $k > 1$. The statistic we would be led to is the V statistic

$$T(\hat{F}_n) = \frac{1}{n^k} \sum_{1 \leq i_1, \ldots, i_k \leq n} \psi(X_{i_1}, \ldots, X_{i_k}). \tag{2.14}$$

Typically, however, one considers the equivalent

$$U_n \equiv T\left(\hat{F}_n, \frac{1}{n}\right) = \frac{1}{\binom{n}{k}} \sum_{1 \leq i_1 < \ldots < i_k \leq n} \psi(X_{i_1}, \ldots, X_{i_k}) \tag{2.15}$$

where $T(F, s) = \prod_{i=1}^{k-1} \frac{1}{(1-is)} \int \ldots \int_{x_1 < \ldots < x_k} \psi(x_1, \ldots, x_k) \prod_{j=1}^{k} dF(x_j)$ and $T(F, 0) = T(F)$ if F is continuous.

This example is not quite covered by our theory on two grounds.

(i) $F \to T(F)$ is not Hadamard differentiable with respect to any of the usual metrics unless ψ is of bounded variation.

(ii) $T(\hat{F}_n)$ is not the statistic U_n one wants to consider.

Both are covered by noting that all we need to do for (ii) is to replace $T(F)$ by $T(F, s)$, $0 \leq s \leq 1$ and $\mathcal{F}$ by $\mathcal{F} \times [0, 1]$, following a suggestion of Reeds (1976). For (i) we note that (2.7) and (2.8) can be established directly for such statistics, (Arcones and Giné, 1993). So again, bootstrapping $\sqrt{n}\left(T\left(\hat{F}_n^*, 1/n\right) - T\left(\hat{F}_n, 1/n\right)\right)$ gives the correct answer.

Another interesting possibility suggested by this example is $\dot{T}(F) = 0$ in which case the limit law $\mathcal{L}_0$ is point mass at 0. We need to renormalize. It is easy to see (Bretagnolle, 1981) that in this case, (2.7) holds with n replacing $\sqrt{n}$ and a suitable limit, but (2.8) fails. It is possible to bring this example also into our framework obtaining a solution proposed by Arcones and Giné (1993), but the hypothesis implicitly tested, $H_0 : T(F) = 0$, $\dot{T}(F) = 0$ is somewhat artificial. □

Next is a complex example illustrating the broad applicability of our approach to semiparametric hypotheses.

Example 6 *Test of change-point* (Matthews, Farewell and Pyke, 1985). Consider a parametric problem where F has the following hazard rate function:

$$\lambda(t) = \begin{cases} \lambda, & \text{if } 0 \le t < \theta \\ (1-\xi)\lambda, & \text{if } t \ge \theta \end{cases} \tag{2.16}$$

where $0 \le \xi < 1$ and $\lambda > 0$ are unknown, and $\theta \ge 0$ is the unknown change-point parameter for the hazard rate which changes from λ to $(1-\xi)\lambda$ at time θ. If θ is confined to a finite interval $[\theta_1, \theta_2]$, the following test statistic was proposed on maximum likelihood grounds by Matthews, Farewell and Pyke (1985) for the irregular hypothesis $H_0 : \xi = 0$ vs. $H_1 : \xi \neq 0$

$$T_n = \sup_{\theta_1 \le \theta \le \theta_2} |\hat{Z}_n(\theta)| \tag{2.17}$$

where $\hat{\lambda}_n = \lambda(\hat{F}_n) = \left(\int x \, d\hat{F}_n(x)\right)^{-1}$ and

$$\hat{Z}_n(\theta) = (1 - e^{-\hat{\lambda}_n \theta})^{-1/2} (n e^{\hat{\lambda}_n \theta})^{1/2} \int_\theta^\infty ((x-\theta)\hat{\lambda}_n - 1) \, d\hat{F}_n(x). \tag{2.18}$$

Under H_0 we have $F(x) = F_\lambda(x) = 1 - e^{-\lambda x}$, thus by integration by parts, $\hat{Z}_n$ can be expressed as

$$\hat{Z}_n(\theta) = -(1-e^{-\hat{\lambda}_n \theta})^{-1/2} e^{\hat{\lambda}_n \theta/2} \sqrt{n} \left\{ \hat{\lambda}_n \int_\theta^\infty U(x, \hat{F}_n) dx - U(\theta, \hat{F}_n) \right\}, \tag{2.19}$$

where $U(x, \hat{F}_n) = \hat{F}_n(x) - F_{\lambda(\hat{F}_n)}(x)$. The limiting distribution of T_n in (2.19) is studied by Matthews, Farewell and Pyke (1985).

Now suppose that we have right censored data or doubly censored data as in Example 4. Then the analogous test statistic is obtained by making $\hat{F}_n$ in $U(\cdot, \hat{F}_n)$ and $\hat{\lambda}_n = \lambda(\hat{F}_n)$ be the NPMLE (the Kaplan–Meier estimate for right censored data) of F for doubly censored data. We modify T_n slightly

to avoid the usual technicalities in censoring replacing $\hat{Z}_n$ by $\hat{Z}_n^M$ given by

$$\hat{Z}_n^M(\theta) = -(1-e^{-\hat{\lambda}_n\theta})^{-1/2}e^{\hat{\lambda}_n\theta/2}\sqrt{n}\left\{\hat{\lambda}_n\int_\theta^M U(x,\hat{F}_n)dx - U(\theta,\hat{F}_n)\right\} \tag{2.20}$$

where $M > \theta_2$. Under H_0 and some regularity conditions, $\hat{\lambda}_n$ converges to λ in probability and $\sqrt{n}U(\cdot,\hat{F}_n)$ converges weakly to a centered Gaussian process G_λ on $[\theta_1, M]$. Therefore we know that under H_0, $\hat{Z}_n^M(\theta)$ weakly converges to

$$Z^M(\theta) = -(1-e^{-\lambda\theta})^{-1/2}e^{\lambda\theta/2}\left\{\lambda\int_\theta^M G_\lambda(x)dx - G_\lambda(\theta)\right\} \tag{2.21}$$

and

$$T_n = \sup_{\theta_1\le\theta\le\theta_2}|\hat{Z}_n^M(\theta)| \xrightarrow{D} \sup_{\theta_1\le\theta\le\theta_2}|Z^M(\theta)|, \text{ as } n\to\infty. \tag{2.22}$$

Write $\hat{Z}_n^M(\cdot) = \sqrt{n}\,T(\hat{F}_n)$, where

$$T(F)(\theta) = -(1-e^{-\lambda(F)\theta})^{-1/2}e^{\lambda(F)\theta/2}\left\{\lambda(F)\int_\theta^M U(x,F)dx - U(\theta,F)\right\}. \tag{2.23}$$

To obtain critical values for T_n, again we simply need to use the bootstrap distribution of $\sup\{\sqrt{n}|T(\hat{F}_n^*)(\theta) - T(\hat{F}_n)(\theta)|;\ \theta_1 \le \theta \le \theta_2\}$. □

Our next example illustrates another extension of the paradigm beyond Theorem 2.1.

Example 7 *Goodness of fit test using kernel density estimates.* Consider the problem of Bickel and Rosenblatt (1973). We have observations on $\mathcal{X} = R$ and wish to test $H_0 : F = F_0$ where $F' = f$ exists and, in fact, $\|f''\|_\infty \le M_0 < \infty$ for all $F \in \mathcal{F}$, and $\inf\{f(x); |x| \le M\} \ge \epsilon > 0$ for some constant $M > 0$.

A natural test statistic is $\sup\{|\hat{f}_n(x) - E_0\hat{f}_n(x)|; |x| \le M\}$, where

$$\hat{f}_n(x) = \int K_{h_n}(x-y)\,d\hat{F}_n(y) \tag{2.24}$$

and $K_h(y) \equiv h^{-1}K(y/h)$, where K is at least twice differentiable, has compact support, is symmetric about 0, $\int_{-\infty}^{\infty}K(y)dy = 1$, and $h_n = n^{-\beta}$, $0 < \beta < 1$ for some β.

This test is consistent against $F \ne F_0$ concentrating on $[-M, M]$, and the natural $T_n(F) = T_n(\cdot, F)$ here is

$$T_n(\cdot, F) = \int K_{h_n}(\cdot - y)\,d(F(y) - F_0(y)).$$

Of course, $\sqrt{n}\,T_n(\hat{F}_n)$ diverges, but

$$U_n(x) \equiv \sqrt{nh_n}\,(T_n(x,\hat{F}_n) - T_n(x,F)) \Rightarrow \mathcal{N}(0,\sigma^2(x,F)), \tag{2.25}$$

where $\sigma^2(x,F) = f(x)\int K^2(y)dy$, and for $x \neq y$, $U_n(x)$ and $U_n(y)$ are asymptotically independent. If our prescription is valid, the distribution of the test statistic

$$T_n = \sup\{|U_n(x)|; |x| \leq M\} \tag{2.26}$$

should be approximable by the bootstrap distribution of

$$\tilde{T}_n^* \equiv \sup\{\sqrt{nh_n}\,|T_n(x,\hat{F}_n^*) - T_n(x,\hat{F}_n)|; |x| \leq M\}. \tag{2.27}$$

This is valid, under the conditions of Bickel and Rosenblatt[4] (1973), by applying the strong approximation to the empirical process and extreme value theory used in Bickel and Rosenblatt (1973) to satisfy the conditions of Theorem 2.2 below.

Suppose $T_n(\hat{F}_n)$, as in Theorem 2.1, are such that

(A1′):

$$\|T_n(\hat{F}_n) - T_n(F) - \dot{T}_n(F)(\hat{F}_n - F)\|_\infty = o_p(n^{-1/2}) \tag{2.28}$$

for all $F \in \mathcal{F}_0$. Suppose further that $\mathcal{X}$ is such that a strong approximation theorem of the following form applies:

(A2′): There exists a probability space on which we can construct $l_\infty(\mathcal{H}_x)$ valued random elements $\sqrt{n}(\hat{\tilde{F}}_n - F)$ having the same distribution as $\sqrt{n}(\hat{F}_n - F)$ and also $l_\infty(\mathcal{H}_x)$ valued Gaussian random elements $\tilde{Z}_F(\cdot)$ with mean 0 and the same covariance structure as $\sqrt{n}(\hat{F}_n - F)$ for which both

$$\|\sqrt{n}(\hat{\tilde{F}}_n - F)(\cdot) - \tilde{Z}_F(\cdot)\|_\infty = o_p(n^{-\gamma}) \tag{2.29}$$

and

$$\|\sqrt{n}(\hat{F}_n^* - \hat{F}_n)(\cdot) - \tilde{Z}_F(\cdot)\|_\infty = o_p(n^{-\gamma}). \tag{2.30}$$

For such results see Csörgő and Revesz (1983), Massart (1989), and Einmahl (1989).

Theorem 2.2 *Under* (A1′) *and* (A2′) *suppose that*

$$\|\dot{T}_n(F)\|_\infty = O(n^\gamma) \tag{2.31}$$

[4]Bickel and Rosenblatt (1973) did not use the Komlos–Maior–Tusnady (1976) strong approximation, so their results and the bootstrap extension are weaker than they need be, which is why the optimal bandwidth $h_n = n^{-1/5}$ is used in our simulation studies in Section 4.

and τ is a seminorm, $\tau(c\boldsymbol{t}) = c\tau(\boldsymbol{t})$ for $c \geq 0$, and τ is subadditive. Suppose also that there exist $\{a_n\}$, $\{b_n\}$ scalar possibly depending on F such that

$$a_n \tau(\dot{T}_n(F) n^{-1/2} Z_F) + b_n \Rightarrow \mathcal{L}_F \tag{2.32}$$

and $a_n = o(n^{1/2})$. Then,

$$a_n \tau(T_n(\hat{F}_n)) + b_n \Rightarrow \mathcal{L}_F$$

and in probability

$$a_n \tau(T_n(\hat{F}_n^*) - T_n(\hat{F}_n)) + b_n \Rightarrow \mathcal{L}_F.$$

Proof By our previous argument and (A1′)

$$T_n(\hat{F}_n^*) - T_n(\hat{F}_n) = \dot{T}_n(F)(\hat{F}_n^* - \hat{F}_n) + o_p(n^{-1/2}) \tag{2.33}$$

and the corresponding statement holds for $T_n(\hat{F}_n) - T_n(F)$.

Under (A2′) and since τ is a seminorm,

$$\begin{aligned} a_n \tau(\dot{T}_n(F)(\hat{F}_n - F)) + b_n &= a_n \tau(\dot{T}_n(F) n^{-1/2} \tilde{Z}_F) \\ &\quad + b_n + O_p(a_n \|\dot{T}_n(F)\|_\infty n^{-(1/2+\gamma)}) \\ &= a_n \tau(\dot{T}_n(F) \tilde{Z}_F n^{-1/2}) + b_n + o_p(1) \Rightarrow \mathcal{L}_F. \end{aligned}$$

The last identity uses (2.31) and $a_n = o(n^{1/2})$. But, under H_0,

$$\begin{aligned} a_n \tau(T_n(\hat{F}_n)) + b_n &= a_n \tau(T_n(\hat{F}_n) - T_n(F)) + b_n \\ &= a_n \tau(\dot{T}_n(F)(\hat{F}_n - F)) + b_n + O_p(a_n n^{-1/2}) \end{aligned}$$

so that $a_n \tau(T_n(\hat{F}_n)) + b_n \Rightarrow \mathcal{L}_F$. The same argument applies to $a_n \tau(T_n(\hat{F}_n^*) - T_n(\hat{F}_n)) + b_n$ and the theorem follows. ∎

We close this section with an old example in which although our formalism applies, the conditions (A1) or (A1′) of our theorems fail and our solution is incorrect.

Example 8 *Test of distribution support.* Suppose $\mathcal{F} = \{F; F$ has support on $[0, b]$ with unknown b, continuous density f and $f(b-) > 0\}$. Then, as is well known, if $X_{(1)} < \ldots < X_{(n)}$ are the ordered X_i's, $n(b - X_{(n)})$ has a limiting distribution $(f(b-))^{-1}\mathrm{Exp}(1)$, where $\mathrm{Exp}(\mu)$ denotes the exponential distribution with mean μ. Thus the natural test statistic for $H_0 : b = b_0$ is $T_n = n(X_{(n)} - b_0)$. If we let $T(F) = F^{-1}(1) - b_0$, we have put the hypothesis in our framework and have noted that under H_0,

$$-T_n = -nT(\hat{F}_n) \Rightarrow (f(b_0-))^{-1}\mathrm{Exp}(1).$$

However, the bootstrap distribution of $n(T(\hat{F}_n^*) - T(\hat{F}_n))$ does not converge as was already noted by Bickel and Freedman (1981) — see also BGvZ. Although Putter and van Zwet (1996) gave a method for repairing bootstrap inconsistency for a similar case in their Example 3.2, there is a much more general solution for this problem discussed in BGvZ, which we recapitulate and discuss briefly in the next section. □

3 The m out of n bootstrap hypothesis tests

This method, presented generally in Bickel, Götze and van Zwet (1997) (BGvZ) and in an alternative form by Politis and Romano (1996), is based on the assumption that under $H_0 : F \in \mathcal{F}_0$, the test statistic, $T_n = T_n(\hat{F}_n)$ is such that

$$T_n \Rightarrow \mathcal{L}_F \tag{3.1}$$

which is nondegenerate. The m out of n bootstrap prescribes that the appropriate quantile of the bootstrap distribution of $T_m(\hat{F}_m^*)$ be used, that is of the distribution of the statistic based on m observations resampled from $X_1, \ldots, X_n$. The history of this approach which goes back to Bickel and Freedman (1981) and Bretagnolle (1981) is partially reviewed in BGvZ. If $m \to \infty$ and $\frac{m}{n} \to 0$ the prescription succeeds in giving an asymptotically correct level under very mild conditions which we detail below. Politis and Romano (1996) argue that by resampling without replacement this conclusion holds with no conditions.

Bickel and Ren (1996) checked the regularity condition for the applicability of this method in Example 4 when data are doubly censored. BGvZ shows its applicability in Example 8. Here is a formal theorem. Let T_n be as above, $T_m^* \equiv T_m(\hat{F}_m^*)$ and C_α^* be given by

$$P_n\{T_m^* \geq C_\alpha^*\} = \alpha. \tag{3.2}$$

Let $\mathcal{H} = \{h : \mathbb{R} \to \mathbb{R};\ |h(x) - h(y)| \leq |x - y|, \|h\| \leq 1\}$, and for $h \in \mathcal{H}$,

$$\theta_m(F) \equiv E_F\{h(T_m(X_1, \ldots, X_m; F))\},$$

where $T_m(X_1, \ldots, X_m; F) \equiv T_m(X_1, \ldots, X_m) - \mu_m(F)$ and $\mu_m(F) = 0$ if $F \in \mathcal{F}_0$. Furthermore, with $X_i^{(j)} \equiv (X_i, \ldots, X_i)_{1\times j}$,

$$\theta_{m,n}(F) \equiv \frac{1}{n^m} \sum_{r=1}^{m} \sum_{i_1+\ldots+i_r=m, i_j \geq 0} \binom{n}{r} \binom{m}{i_1, \ldots, i_r} E_F\{h(T_m(X_1^{(i_1)}, \ldots, X_r^{(i_r)}; F))\}.$$

Theorem 3.1 *Let $m = o(n)$ with $m \to \infty$, as $n \to \infty$ and let C_α^* be given by* (3.2). *Assume for $0 < \alpha < 1$*

(a) $T_n(X_1, \ldots, X_n; F) \Rightarrow W_F$ *where* $W_F \sim \mathcal{L}_F$ *for all* $F \in \mathcal{F}$ *and* $\mathcal{L}_F$ *is continuous for* $F \in \mathcal{F}_0$.

(b) $\mu_n(F) \to \infty$, $\mu_m(F) - \mu_n(F) \to -\infty$ *uniformly on bounded Lipschitz compacts contained in* $\mathcal{F}_0^c$ *if* $m \to \infty$, $\frac{m}{n} \to 0$.

(c) $\sup_{h \in \mathcal{H}} |\theta_{m,n}(F) - \theta_m(F)| = o(1)$, *for all* $F \in \mathcal{F}$.

Then,

(i) $\lim_{n\to\infty} P\{T_n \geq C_\alpha^* \mid H_0\} = P\{W_F \geq C_\alpha^0 \mid F \in \mathcal{F}_0\} = \alpha$;

(ii) For alternatives $H_n : F = F_n$ *such that for some* $F_0 \in \mathcal{F}_0$, $\{F_n\}$ *are contiguous to* F_0, *we have that under* H_n,

$$C_\alpha^* \xrightarrow{P} C_\alpha^0, \text{ as } n \to \infty$$

and hence the tests based on the critical values C_α^* *and* C_α^0 *have the same asymptotic power functions;*

(iii) For a fixed alternative $H_1 : F = F_1 \notin \mathcal{F}_0, P\{T_n \geq C_\alpha^* \mid H_1\} \to 1$, *as* $n \to \infty$.

Remark. Assumption (c) essentially says that T_m is not really perturbed by $o(\sqrt{n})$ ties in its arguments — see BGvZ.

Proof Theorem 2 in BGvZ shows that (c), $m \to \infty$, $m/n \to 0$ implies that,

$$E^* h(T_m^*) \xrightarrow{P} E_F h(W_F) \tag{3.3}$$

where $W_F \sim \mathcal{L}_F$. But bounded Lipschitz convergence is equivalent to weak convergence. Thus, noting that under H_0, the identity $T_n(X_1, \ldots, X_n; F) = T_n(X_1, \ldots, X_n)$, (3.3) and Polya's theorem imply that $C_\alpha^* \xrightarrow{P} C_\alpha^0$ and another application of Polya's theorem yields (i). Assertion (ii) follows from the definition of contiguity. To argue for (iii), note that Theorem 2 of BGvZ implies that for all F

$$T_m^* - \mu_m(\hat{F}_n) \Rightarrow \mathcal{L}_F. \tag{3.4}$$

Therefore under $F \notin \mathcal{F}_0$,

$$C_\alpha^* - \mu_m(\hat{F}_n) = O_p(1). \tag{3.5}$$

But $T_n - \mu_m(\hat{F}_n) = T_n - \mu_n(\hat{F}_n) + (\mu_n(\hat{F}_n) - \mu_m(\hat{F}_n)) \xrightarrow{P} \infty$. Then to reject iff $T_n \geq C_\alpha^*$ is equivalent to reject iff $T_n - \mu_m(\hat{F}_n) \geq C_\alpha^* - \mu_m(\hat{F}_n)$. The result follows from (3.5). ∎

The proof shows that $C_\alpha^* \xrightarrow{P} \infty$ if $F \notin \mathcal{F}_0$ and thus we expect that the power of this test is less than that of the tests proposed in Section 2 where

these are valid. We give some simulations to show that this is indeed the case. The question naturally presents itself: Is there a way of correcting the m out of n bootstrap to give results comparable to those we obtain by simulating the tests of Section 2? A systematic answer is given in Bickel and Sakov (1999) (in preparation) and the 1998 thesis of Sakov.

We note that the m out of n bootstrap has the additional advantage of computational savings, see Bickel and Yahav (1988) for instance. In fact the computational savings can be garnered in the context of Section 2 also. Specifically it is clear that the conclusions of Theorem 2.1 continue to hold if the bootstrap distribution of $\tau(\sqrt{n}(T(\hat{F}_n^*)-T(\hat{F}_n)))$ is replaced in calculating the critical value by that of $\tau(\sqrt{m}(T(\hat{F}_m^*)-T(\hat{F}_n)))$ as long as $m \to \infty$. It is intuitively clear that $m << n$ may give poor critical values. But, in practice, the effect as long as m is moderate seems small in the simulations we have conducted. Further investigation is necessary.

4 Simulations

In this section we present some simulation results exhibiting the success of the method given in Theorem 2.1 and Corollary 2.1 by a number of our examples and the inferior behavior of the m out of n bootstrap in all cases but Example 8.

We give simulations for Example 3 — the median test, Example 4 — the goodness of fit test with doubly censored data, Example 7 and Example 8. In our studies, the following power curves are compared:

$$P_0(\theta) = P\{T_n \geq C_\alpha^{(n)} | \theta\}, \qquad P_0 : \text{————}$$

$$P_n(\theta) = P\{T_n \geq C_\alpha^{*(n)} | \theta\}, \qquad P_n : \text{- - - - - - -}$$

$$P_m(\theta) = P\{T_n \geq C_\alpha^{*} | \theta\}, \qquad P_m : \cdots\cdots\cdots$$

where $\alpha = 0.05$, T_n is the test statistic, $C_\alpha^{(n)}$ is the true critical value obtained by the Monte Carlo method, $C_\alpha^{*(n)}$ is the critical value based on the adjusted n out of n bootstrap as in Corollary 2.1, C_α^* is the critical value based on the m out of n bootstrap as in Theorem 3.1, and θ is the parameter used to compute the power of the test. For each simulation run, $C_\alpha^{*(n)}$ and C_α^* are based on 400 bootstrap samples, and $P_0(\theta), P_n(\theta), P_m(\theta)$ are computed based on 400 random samples for each θ.

(I). In Example 3, we consider test $H_0 : \theta = 0$ vs. $H_1 : \theta > 0$, where θ is the median of the distribution F from which $X_1, \ldots, X_n$ is drawn. Figure 1 compares the power curves P_0, P_n and P_m, where $n = 400$, F is the normal distribution with mean θ and variance 25, and all power curves are the average of 500 simulation runs.

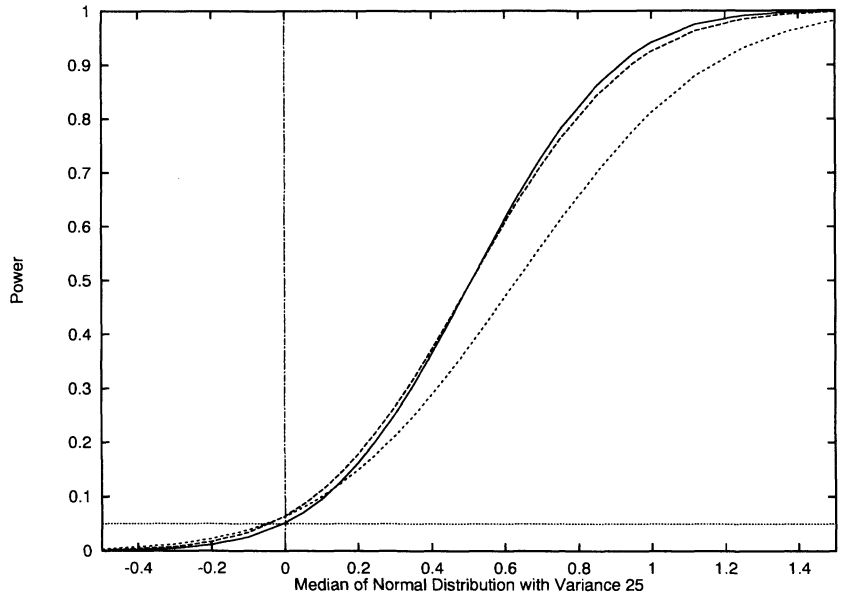

Figure 1. Power curves of median test with complete sample of size 400.

Here $m = \sqrt{n} = 20$ is used since for the median an Edgworth expansion to terms of order $\frac{1}{\sqrt{n}}$ is valid if the density F has a finite derivative F' and the Edgworth expansion for the m out of n bootstrap is valid for $m = O(\sqrt{n})$ under the same conditions with the same leading term and error terms $\frac{1}{\sqrt{m}}$ and $\sqrt{\frac{m}{n}}$ (Sakov and Bickel, 1998). The "optimal" rate of m then balances $\sqrt{\frac{m}{n}}$ and $\frac{1}{\sqrt{m}}$ to give $m = \sqrt{n}$.

(II). In Example 4, we consider the goodness of fit test $H_0 : G = G_0$ vs. $H_1 : G \neq G_0$ for doubly censored data using the Cramér-von Mises test statistic:

$$T_n = n \int (\hat{G}_n - G_0)^2 dG_0.$$

Denoting $\text{Exp}(\theta)$ as the exponential distribution with mean θ, for $n = 200, m = \sqrt{n}, G_0 = \text{Exp}(1), C = \text{Exp}(3), B = \frac{2}{3}C - 2.5$ (which, under H_0, gives 55.7% uncensored, 25.2% right censored and 19.1% left censored observations), Figure 2 compares the power curves P_0, P_n and P_m, which are the average of 100 simulation runs.

(III). In Example 7, we consider test $H_0 : F = F_0$ vs. $H_1 : F \neq F_0$, with $F_0 = \text{Exp}(1)$. For test statistic T_n given by (2.26) with $n = 400, h_n = n^{-1/5}, M = 3, K = U(-1, 1)$ and θ as the mean of the exponential distribution, Figure 3 compares the power curves P_0, P_n and P_m, which are the average of 100 simulation runs. Here for $m = \sqrt{n}$, the power curve P_m by the m out of n bootstrap uses the critical value based on $\hat{T}_m^* =$

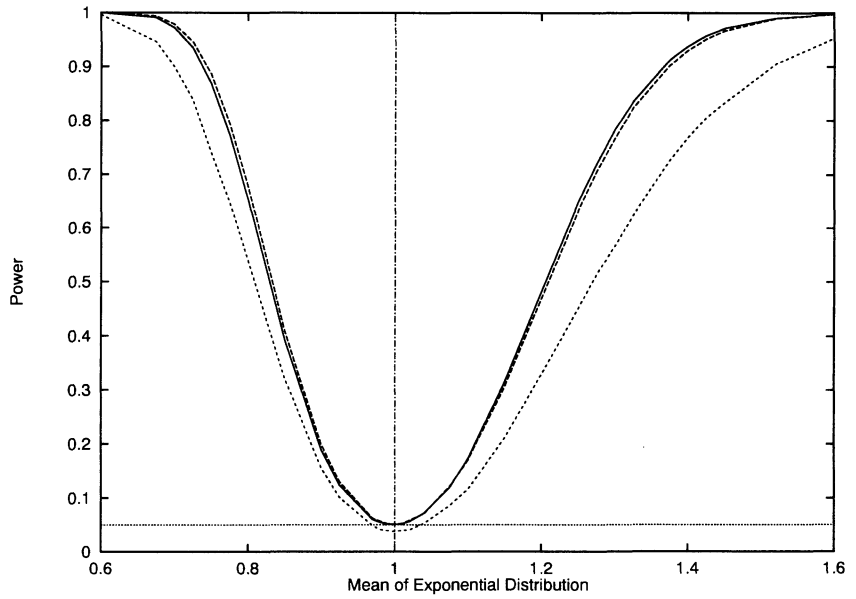

Figure 2. Power Curves of GOF Test with Doubly Censored Sample of Size 200.

$\sqrt{mh_m}\sup\{|\hat{f}_m^* - \hat{f}_n|; |x| \leq M\}$, which coincides with $\tilde{T}_n^*$ given by (2.27) if $m = n$.

(IV). In Example 8, we consider test $H_0 : b = 1$ vs. $H_1 : b > 1$, with $F = U(0, b)$. For $n = 400, m = n^{1/3}$ and $\theta = b$, Figure 4 compares the power curves P_0, P_n and P_m, which are the average of 1000 simulation runs.

In this case, the power function $P_n(\theta)$, when the adjusted n out of n bootstrap is used, is a total breakdown under H_0. One should note that in Figure 4, the power $P_n(\theta)$ under H_0, i.e., when $\theta = 1$, is always 0, while $\alpha = 0.05$, although it seems that $P_n(\theta)$ and $P_0(\theta)$ are quite close overall. Here the heuristics based on the asymptotic expansion the distribution of the maximum whose first error term is of order $\frac{1}{n}$ and heuristics discussed in BGvZ suggest that an appropriate order of $m = n^{1/3}$, in this case $m = 7$.

5 Appendix

We give brief arguments for the validity of the application of Theorem 2.1 in our examples.

Example 1 Taking $\mathcal{H}_x = \mathcal{H}_y =$ indicators of rays for R and τ corresponding to the Kolmogorov, Smirnov and Cramér - von Mises tests are covered by Corollary 2.1 as are the analogous tests when one takes $\mathcal{H}_x$ to be a universal Donsker class in higher dimensions (van der Vaart and Wellner (1996).

Example 2 Suppose the model $\mathcal{F}$ is regular and $\theta(\hat{F}_n)$ is a regular estimate in the sense of BKRW. Suppose also that $\theta : \mathcal{F} \to R^d$ is Hadamard differen-

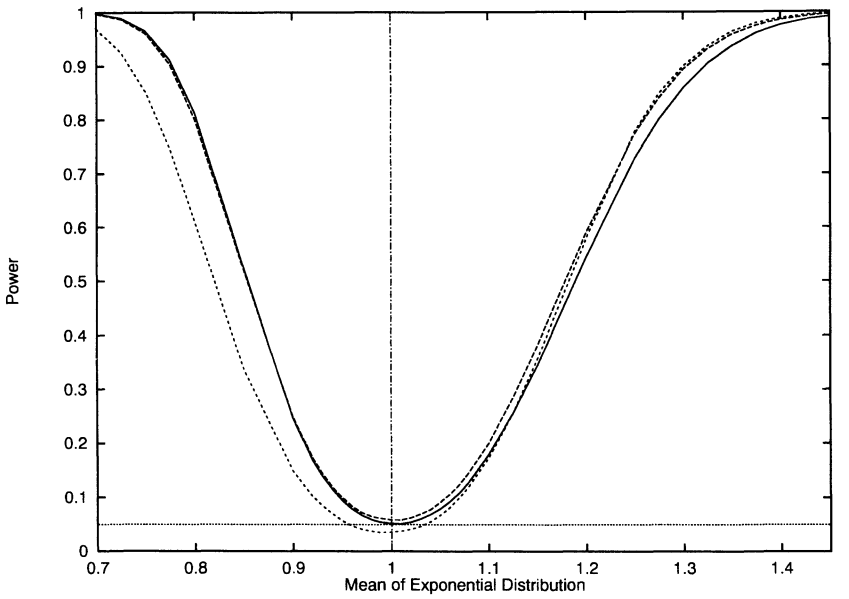

Figure 3. Power Curves of GOF Test Using Density with Complete Sample of Size 400.

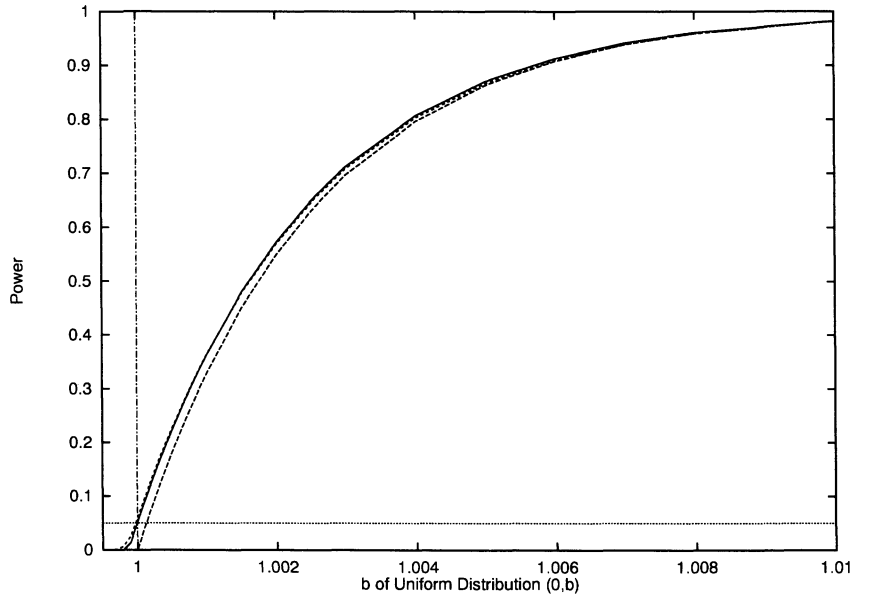

Figure 4. Power Curves of Support Test with Complete Sample of Size 400.

tiable with respect to $\|\cdot\|_\infty$ (in $l_\infty(\mathcal{H}_x)$) with derivative $\dot{\theta} : \bar{\mathcal{F}} \to R^d$. Then $F \to F_{\theta(F)}$ is Hadamard differentiable since $\theta \to F_\theta$ is Hadamard differentiable from R^d to $\mathcal{F}_0$ by the regularity of the model and thus the composition $F \to \theta(F) \to F_{\theta(F)}$ is also.

Example 3 The satisfaction of the conditions here on the sets $\mathcal{F} = \{F : E_F|X|^{2+\delta} < \infty,\ \delta > 0\}$ and $\mathcal{F} = \{F : f' > 0\}$ is well known.

Example 4 The appropriateness of the conditions for right censored data may be obtained from Anderson, Borgen, Gill and Keiding (ABGK) (1993) and for the doubly censored case in Bickel and Ren (1996).

Example 5 Appropriate references are cited in the example.

Example 6 The arguments for the uncensored case is in Matthews et al (1985). The censored case modifications are clear from the theory of the Kaplan–Meier for right censored data or the NPMLE for doubly censored data (see Bickel and Ren, 1996).

Example 7 The arguments based on Bickel and Rosenblatt (1973) are sketched in the example.

Example 8 The arguments are given in BGvZ.

REFERENCES

Anderson, P. K., Borgan O., ϕ, Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer–Verlag.

Arcones, M. and Gine, E. (1993). Limit theorems for U processes. *Annals of Probability* 4, 1449–1452.

Beran, R. (1986). Simulated power functions. *Ann. Statist.* 14, 151–173.

Beran, R. and Millar, P. W. (1987). Stochastic estimation and testing. *Ann. Statist.* 15, 1131–1154.

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* Vol. 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.* (To appear).

Bickel, P. J., Klaassen, C., Ritov, Y. and Wellner, J. (1993,1998). *Efficient and Adaptive Estimation in Semiparametric Models.* Johns Hopkins Press, Baltimore, Springer, New York.

Bickel, P. J. and Ren, J. (1996). The m out of n bootstrap and goodness of fit tests with doubly censored data. *Robust Statistics, Data Analysis, and Computer Intensive Methods.* Lecture Notes in Statistics. Springer–Verlag, 35–47.

Bickel, P. J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *An. Statist.* 1, 1071–1095.

Bickel, P. J. and Sakov, A. (1998). On the choice of m in the m out of n bootstrap.(Preprint)

Bickel, P. J. and Yahav, J. A. (1988). Richardson extrapolation and the bootstrap. *J. Amer. Statist. Assoc.* 83, 387–393.

Bretagnolle, J. (1981). Lois limites du bootstrap de certaines fonctionelles. *Ann. Inst. H. Poincaré*, Ser. B 19, 281–296.

Csörgő, M. and Revesz, P. (1981). *Strong Approximations in Probability and Statistics.* Academic Press. New York.

Freedman, D. A. (1981). Bootstrapping regression models. *Ann. Statist.* 12, 1218–1228.

Galambos, J. (1978). *The Asymptotic Theory of Extreme Order Statistics.* John Wiley & Sons, New York.

Gehan, E. A. (1965). A generalized two-sample Wilcoxon test for doubly censored data. *Biometrika* 52, 650–653.

Gill, R. D. (1983). Large sample behavior of the product-limit estimator on the whole line. *Ann. Statist.* 11, 49–58.

Giné, E. and Zinn, J. (1990). Bootstrapping general empirical measures. *Ann. Prob.* 18, 852–869.

Gu, M. G. and Zhang, C. H. (1993). Asymptotic properties of self-consistent estimators based on doubly censored data. *Ann. Statist.* Vol. 21, No. 2, 611–624.

Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.* 21, 1926–1947.

Hawkins, D. L., Kochar, S. and Loader, C. (1992). Testing exponentially against IDMRL distributions with unknown change point. *Ann. Statist.* Vol. 20, 280–290.

Hinkley, D. V. (1987). Bootstrap significance tests. *Proceedings of the 47th Session of International Statistical Institute*, Paris, 65–74.

Hinkley, D. V. (1988). Bootstrap methods. *J. R. Statist. Soc.* B, 50, 321–337.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* 53, 457–481.

Komlós, J., Maior, P. and Tusnády, K. (1976). An approximation of partial sums of independent r.v.s. and the sample df. *Z. Wahrsch. Verw. Gebiete* 32, 33–58.

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

Massart, P. (1989). Strong approximations for the multivariate empirical and related processes by KMT construction. *Ann. Probab.* 17, 266–291.

Matthews, D. E., Farewell, V. T. and Pyke, R. (1985). Asymptotic score-statistic processes and tests for constant hazard against a change-point alternative. *Ann. Statist.* Vol. 13, 583–591.

Mykland, P. and Ren, J. (1996). Algorithms for computing the self-consistent and maximum likelihood estimators with doubly censored data. *Ann. Statist.* 24, 1740–1764.

Politis, D. N. and Romano, J. P. (1996). A general theory for large sample confidence regions based on subsamples under minimal assumptions. *Ann. Statist.* 22, 2031–2050.

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

Rao, C. R. (1973). *Linear Statistical Inference and its Applications.* Wiley, New York.

Reeds, J. A. (1976). *On the Definition of von Mises Functionals.* Ph.D. Dissertation, Harvard University, Cambridge, Massachusetts.

Ren, J. (1995). Generalized Cramér–von Mises tests of goodness of fit for doubly censored data. *Ann. Inst. Statist. Math.* 47, 525–549.

Romano, J. P. (1988). A bootstrap revival of some nonparametric distance tests. *J. Amer. Statist. Assoc.* 83, 698–708.

Romano, J. P. (1989). Bootstrap and randomization tests of some nonparametric hypotheses. *Ann. Statist.* 17, 141–159.

Sakov, A. (1998). Ph.D. Thesis, University of California - Berkeley.

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap.* Springer, New York.

Shorack, G. R.and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics.* John Wiley & Sons, Inc.

Turnbull, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *J. Amer. Statist. Assoc.* 69, 169–173.

Vaisman, I. (1984). *A First Course in Differential Geometry.* Marcel Dekker, Inc., New York.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes.* Springer, New York.

PETER J. BICKEL
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CA 94720
USA
bickel@stat.Berkeley.edu

JIAN-JIAN REN
DEPARTMENT OF MATHEMATICS
TULANE UNIVERSITY
NEW ORLEANS, LA 70118
USA
renj@ultra1.math.tulane.edu