On the Choice of m in the m Out of n Bootstrap and its Application to Confidence Bounds for Extreme Percentiles *†

Peter J. Bickel
University of California, Berkeley ‡
Anat Sakov
Tel-Aviv University

June 27, 2005

Abstract

The m out of n bootstrap Bickel et al. [1997], Politis and Romano [1994] is a modification of the ordinary bootstrap which can rectify bootstrap failure when the bootstrap sample size is n. The modification is to take bootstrap samples of size m where $m \to \infty$ and $m/n \to 0$. The choice of m is an important matter, in general. In this paper we consider an adaptive rule proposed by Bickel, Götze and van Zwet (personal communication) to pick m. We give general sufficient conditions for validity of the rule and then examine its behavior in

AMS 1991 *subject classification.* Primary 62G09; secondary 62G20, 62G30.

†*Key words and phrases.* bootstrap, m out of n bootstrap, choice of m, data-dependent rule, adaptive choice.

‡Supported by NSF Grant DMS-98-02960

the problem of setting confidence bounds on high percentiles such as the expectation of the maximum. We also study largely in this context optimality properties of this rule paralleling work in Götze and Račkauskas [2001] for smooth functionals. Finally, we consider the rule when the ordinary bootstrap is valid. Simulations complete our results.

1 Introduction

In the last ten years attention has been given to ways of remedying extreme failures (lack of consistency) of Efron's nonparametric bootstrap Efron [1979]. Bickel et al. [1997] gave a catalogue of old and new examples and some guidelines for bootstrap failure. Nevertheless, it is hard to categorize and identify a-priori whether the bootstrap works or not in a specific situation. Bickel et al. [1997],Götze [1993],Politis and Romano [1994] revived a discussion of resampling smaller bootstrap samples, namely, instead of resampling bootstrap samples of size n (referred to here as the n-bootstrap), take bootstrap samples of size m (referred to here as the m-bootstrap), where $m \to \infty$ and $m/n \to 0$ with or without replacement Bretagnolle [1983],Swanepoel [1986]. The choice of m can be crucial, and there are two issues involved: the first is that the user does not know a-priori whether the bootstrap works or not in his particular case. The second problem is the choice of m, if the n-bootstrap fails.

Earlier papers discussing the choice of m include Datta and McCormick [1995] (in the framework of an AR(1) model, and using a jackknife choosing m which minimizes some risk function), and Hall et al. [1995] for dependent data (by balancing between bias and variance). See Politis et al. [1999] for more references and discussion.

Bickel, Götze and van Zwet proposed a data dependent rule for selecting m in estimating the limit laws of pivotal statistics. The study of this rule was independently begun by Götze and Račkauskas [2001] and Sakov [1998]. Both sets of authors gave conditions for the rule to select m such that $m/n \to 0$ and $m \to \infty$ when the bootstrap is inconsistent ("does not work").

Götze and Račkauskas using delicate analysis of the behavior of U-statistics with kernels growing in n focused on conditions under which the rule gives "optimal" choices of m, that is, resulting in an approximation of the same order as the one that could be obtained by an oracle knowing the underlying distribution. They verified their conditions in a number of classical examples such as pivots based on the minimum

of $U(0,\theta)$ variables and quadratic statistics both finite and infinite dimensional. Sakov's focus emphasized quadratic statistics also, but was much less general. On the other hand, she showed how the m out of n bootstrap could, in some situations, be modified using extrapolation ideas to give approximation as good as the best other available approximation, an impossibility even for the oracle Sakov and Bickel [2000] and Bickel and Sakov [2002b].

In this paper we begin with a formulation for what is meant by failure of the bootstrap, in terms of convergence of measure valued random elements which is more abstract than that of Götze and Račkauskas, but we believe clarifies the rationale of our rule. We then state and prove some elementary results on the behavior of our rule, under conditions which are plausible but whose validation can be quite non-trivial, as might be expected from the verifications in Götze and Račkauskas.

The major result of our paper is, in fact, the validation of these conditions for a class of pivotal statistics based on $\max(X_1, ..., X_n)$ for setting a confidence bounds on $F^{-1}(1-1/n)$.

Subsampling, or the m out of n bootstrap without replacement is also an option for this confidence bound problem Politis et al. [1999]. We conjecture that it work just as well. In special cases such as the maximum for the uniform distribution, the $\hat{m}$ chosen is such that $\hat{m}/\sqrt{n} \to 0$ and hence the sampling with and without replacement bootstrap are the same with probability tending to 1. It also seems clear that our results can be generalized to bounds for $F^{-1}(1-c/n)$ and under weaker conditions than the specialized von-Mises assumptions we make. We have chosen not to prove these extensions in this paper.

The paper is organized as follows: in Section 2, we motivate and describe a rule for choosing m. The properties of the rule are discussed in Section 3: subsection 3.1 give rather general criteria under which the chosen m ($\hat{m}$) behaves properly, i.e., $\hat{m}/n \to 0$ and $\hat{m} \to \infty$ when the bootstrap is inconsistent. In subsection 3.2 we study $\hat{m}$, when the Efron bootstrap works and is optimal in the sense of Beran [1982]. Here, if Edgeworth expansions exist then $\hat{m}/n \to 1$, as

it should. In subsection 3.3 we show that if the n-bootstrap is inconsistent, but Edgeworth or similar expansions are available for the m-bootstrap with m as above, then the rule not only behaves properly but in fact gives essentially the best rates possible for estimation of the limit of the population distributions of the statistics we consider viewed as a population parameter. In some cases, the rule gives the best possible minimax rates for estimation of the limit in the sense given in Bickel et al. [1997]. The application to extrema is presented in Section 4. Section 5 presents some simulations supporting the asymptotics. An Appendix with technical arguments completes the paper.

2 The rule

We start with notations and motivation for the rule proposed by Bickel, Götze and van Zwet (personal communication) for choosing m in estimation problems. The rule also provides a diagnostic for n-bootstrap failure, as discussed in Section 3.2 and demonstrated in Section 5. We assume a known rate of convergence of the statistic to a non-degenerate limiting distribution. The case of unknown rate of convergence will be discussed in connection with our main example.

Assume $X_1, \ldots, X_n$ are iid from a distribution $F \in \mathcal{F}$. For convenience suppose $X_1 \in R^d$. Let $T_n = T_n(X_1, \ldots, X_n, F)$ be a random variable with cdf $L_n(x, F) = P(T_n \leq x)$, such that $L_n \stackrel{\mathcal{L}}{\Rightarrow} L$, where L is a non-degenerate distribution. The goal is to estimate or construct a confidence interval for $\theta_n = \gamma(L_n)$, where γ is a functional. The bootstrap estimates L_n, and this estimate, in turn, is plugged into γ to give an estimate for θ_n. For simplicity, in what follows we suppress the dependence of L_n and L on F.

For any positive integer m, let the bootstrap sample $X_1^*, \ldots, X_m^*$ be a sample drawn from $\hat{F}_n$ (the empirical distribution function) and let the m-bootstrap version of T_n be

$$T_m^* = T_m\left(X_1^*, \ldots, X_m^*, \hat{F}_n\right),$$

with bootstrap distribution

$$L_{m,n}^*(x) \equiv P^*\left(T_m^* \leq x\right) = P\left(T_m^* \leq x \,\middle|\, \hat{F}_n\right).$$

We say that the bootstrap 'works' if $L_{m,n}^*$ converges weakly to L in probability for all $m, n \to \infty$ and in particular for $m = n$. From the results in Bickel et al. [1997] and Politis and Romano [1994] it follows that when the bootstrap does not 'work' then under minimal conditions using a smaller bootstrap sample size rectifies the problem which means that although $L_{n,n}^*$ does not have the correct limiting distribution, $L_{m,n}^*$ with "small" but "not too small" m does. For $m \to \infty$, $m/n \to 0$, sampling may be done with or without replacement. If it is done without replacement (known as subsampling) then $L_{m,n}^* \stackrel{\mathcal{L}}{\Rightarrow} L$ Politis

and Romano [1994]. If the resampling is done with replacement, then $L^*_{m,n} \stackrel{\mathcal{L}}{\Rightarrow} L$, only if T_m is not affected much on the order of $\sqrt{m}$ ties Bickel et al. [1997]. Thus, the subsampling method is more general than the m out of n bootstrap in that one can obtain consistency under minimal assumptions. However, the m out of n bootstrap with replacement has the advantage that it allows for the choice of $m = n$, which is not possible for subsampling. In particular, if the n-bootstrap works and is known to be second order correct for some pivotal roots, the selection rule for m includes the particular case $m/n \to 1$ (See Section 3.2). In that event, the m out of n bootstrap enjoys the second order properties of the n-bootstrap. Since in all situations of interest so far, the conditions for consistency of the bootstrap with replacement are satisfied we consider only that case. The understanding that a smaller bootstrap sample may be needed raises the question of the choice of m.

To motivate the rule we use the following example (for more details see Bickel et al. [1997] and Sakov [1998]). Let $T_n(X_1, \ldots, X_n) = \sqrt{n}\bar{X}_n$, where $X_1, \ldots, X_n$ are iid with unknown mean, μ, and known finite variance σ^2 and we would like to test the null hypothesis that $\mu = 0$. Below we use the subscript on $\bar{X}_n$ to indicate sample size. It is obvious that if one uses the bootstrap in this case he should bootstrap $\sqrt{m}(\bar{X}^*_m - \bar{X}_n)$ in which the bootstrap works in order to find the critical values. However, bootstrapping $\sqrt{m}\bar{X}^*_m$ is a toy example which allows us to discuss the behavior of bootstrap distribution for different m's, as we do below. This example may raise, more naturally, when testing if the expected value of a distribution is 0, and not centering the bootstrap around $\bar{X}_n$, or in regression problem when not centering the residuals Freedman [1981]. Let $T^*_m = \sqrt{m}\bar{X}^*_m$:

1. When m is fixed, say $m = k$ then the bootstrap distribution is

$$L^*_{k,n}(x) = \frac{1}{n^k} \sum_{i_1=1}^{n} \cdots \sum_{i_k=1}^{n} 1\left(\sqrt{k}\frac{\sum_{l=1}^{k} X_{i_l}}{k} \leq x\right),$$

where $1()$ is the indicator function. Note that $L^*_{k,n}$ is a function of the data only and is a V-statistic. It follows that as $n \to \infty$

$L^*_{k,n} \to L_k$ Serfling [1980] and the limit depends on k.

2. When $m, n \to \infty$, $\sqrt{m}(\bar{X}^*_m - \bar{X}_n) \stackrel{\mathcal{L}}{\Rightarrow} \mathcal{N}(0, \sigma^2)$ with probability 1 Bickel and Freedman [1981]. But, this implies that $\sqrt{m}\bar{X}^*_m$ behaves like $\mathcal{N}(\sqrt{m}\bar{X}_n, \sigma^2)$. Now, since $\sqrt{m}\bar{X}_n = \sqrt{\frac{m}{n}}\sqrt{n}\bar{X}_n$, it follows that if $m/n \to \lambda \geq 0$ then under the null hypothesis $\sqrt{m}\bar{X}_n \stackrel{\mathcal{L}}{\Rightarrow} \mathcal{N}(0, \lambda\sigma^2)$, i.e. the limit of $\sqrt{m}\bar{X}^*_m$, denoted by $\mathcal{L}_\lambda$ is the random distribution given by $\mathcal{N}(\sqrt{\lambda}Z, \sigma^2)$ where $Z \sim \mathcal{N}(0, 1)$. Note that $\mathcal{L}_\lambda$ is degenerate and equal to the desired $\mathcal{N}(0, 1)$ iff $\lambda = 0$, i.e. when $m/n \to 0$. However, when $\lambda > 0$, the limit depends on λ. Furthermore, $\lambda \mapsto \mathcal{L}_\lambda$ is $1 - 1$.

Stating loosely what we have just observed: when m is in the "right range" of values the bootstrap distributions are "close" to each other, while when m is too large or too small the bootstrap distributions (or processes) are different. This suggests looking at a sequence of values of m, and their corresponding bootstrap distributions. Using some measure of discrepancy between these distributions will show that the discrepancies are large when m is "too large", they are small when m is the right order, and large again when m fixed since in that case the bootstrap distribution converges to the distribution of the statistic based on m observations for given F.

To state the above more generally: in essentially all examples considered so far 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, in a sense to be made precise in the next section, 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 non-degenerate distribution, $\mathcal{L}_\lambda$, which is typically different from $\mathcal{L}_1$. We expect that $\mathcal{L}_0 = L$.

The above behavior suggests the following rule for choosing m:

1. Consider a sequence of m's of the form

$$m_j = \left[q^j n\right], \qquad \text{for } j = 0, 1, 2, \ldots, \qquad 0 < q < 1, \quad (1)$$

where $[\alpha]$ denotes the smallest integer $\geq \alpha$.

2. For each m_j, find $L^*_{m_j,n}$. In practice this is done by Monte-Carlo.
3. Let $\hat{m} = \underset{m_j}{argmin}\ \rho\left(L^*_{m_j,n}, L^*_{m_{j+1},n}\right)$ where ρ is some metric consistent with convergence in law. If the difference is minimized for a few values of m_j then pick the largest among them. Denote the j corresponding to $\hat{m}$ by $\hat{\jmath}$.
4. The estimator of L is $\hat{L} = L^*_{\hat{m},n}$.
5. Estimate θ by $\hat{\theta}_n = \gamma\left(\hat{L}\right)$ or use the quantiles of $\hat{L}$ to construct confidence interval for θ.

Our discussion and proofs are for the case $\rho(F,G) = \sup_x |F(x) - G(X)|$, the Kolmogorov sup distance. Götze and Račkauskas [2001] consider more general metrics of the form $\rho(P,Q) = \sup\{|P(h)-Q(h)| : h \in \mathcal{H}\}$ where $\mathcal{H}$ is a Donsker class of functions and $P(h) \equiv E_P h(X)$. The results of sections 3.1 and 3.2 are valid for this generalization also, but are not formally pursued. Simulations using other ρ such as the Wasserstein metrics for our application to extrema did not show real differences, although Götze and Račkauskas [2001] obtained better results for $T_n = \sqrt{n}\bar{X}_n$. On the other hand, we do not intend to rule out the use of other metrics in practice. One of us, Sakov, successfully used metrics based on quantiles of P and Q in applications to Silverman's bump test Silverman [1981], Sakov [1998].

3 General behavior of the rule

3.1 Order of $\hat{m}$ when the n-bootstrap is inconsistent

Assume $T_n(X_1, \ldots, X_n, F)$ is the random variable of interest with exact distribution L_n and T_m^* is its bootstrap version with bootstrap distribution $L_{m,n}^*$. Then for a given m the bootstrap distribution is a stochastic process whose distribution depends on $\hat{F}_n$.

To study the behavior of such objects carefully we introduce the following framework. On a single large probability space $(\Omega, \mathcal{A}, P)$ we suppose that we can define:

a) $X_1, \ldots, X_n, \ldots$ iid F on R^d.

b) X_{jk}^* $j \geq 1$, $k \geq 1$ such that the conditional distribution of $X_{1n}^*, X_{2n}^*, \ldots$ given $(X_1, \ldots, X_n)$ is that of iid variables with common distribution $\hat{F}_n$.

We represent the laws of random variables by their distribution functions viewed as elements of the Skorokhod space $D(\bar{R})$. Thus, $L_{m,n}^*$ is a measurable map from Ω to $D(\bar{R})$. By saying that $L_{m,n}^*$ *converges in law to (a random) L in probability* we shall mean: There exist,

(i) Maps $\tilde{L}_{m,n}^* : \Omega \to D(\bar{R})$, $m, n \geq 1$.

(ii) A map $L : \Omega \to C(\bar{R})$, such that

(a) The distributions of $\{L_{m,n}^*\}_{m \geq 1}$ and $\{\tilde{L}_{m,n}^*\}_{m \geq 1}$ agree for all n. That is, for any $k, j_1, \ldots, j_k$

$$P\left(\left(L_{j_1,n}^*, \ldots, L_{j_k,n}^*\right)^{-1}\right)(\cdot) = P\left(\left(\tilde{L}_{j_1,n}^*, \ldots, \tilde{L}_{j_k,n}^*\right)^{-1}\right)(\cdot),$$

or

(a') $\|L_{m,n}^* - \tilde{L}_{m,n}^*\|_\infty = o_p(1)$ as $m, n \to \infty$ and

(b) If ρ is the Skorokhod (or Prohorov) metric,

$$\rho\left(\tilde{L}_{m,n}^*, L\right) \xrightarrow{p} 0, \qquad \text{as } m, n \to \infty. \tag{2}$$

Note: It is possible to combine (a) and (a') into a single condition but with no gain in simplicity.

Since L is continuous with probability 1, (2) implies that

$$\left|\left|\tilde{L}^*_{m,n} - L\right|\right|_\infty \xrightarrow{p} 0. \tag{3}$$

We shall write $L^*_{m,n} \overset{\mathcal{L}}{\Rightarrow} L$ when (2) holds. We will use an extension of these notions by considering $T_n(\cdot)$ whose values themselves lie in R^k, or generally a separable metric function space $(\mathcal{F}, d)$. For instance, consider $T_n\left(\hat{F}_n, F\right) \equiv \sqrt{n}\left(\hat{F}_n - F\right)$. Then by the law, $\mathcal{L}^*\left(T_m\left(\hat{F}^*_m, \hat{F}_n\right)\right)$, we will mean a measurable map from Ω to the space of probability distributions on $D(\bar{R})$ endowed with the Prohorov metric. Now L will similarly be a measurable map from Ω to the space of all probabilities on $C(\bar{R})$. Definition (2) for convergence in law of $\mathcal{L}^*\left(T_m\left(\hat{F}^*_m, \hat{F}_n\right)\right)$ in probability carries over, save that ρ is replaced by the Prohorov metric, and (3) is no longer relevant. In principle, we can consider $T_n(\cdot,\cdot)$ taking values in $l^\infty(\mathcal{F})$ where $\mathcal{F}$ is a set of functions on R^d and formulate results as in van der Vaart and Wellner [1996] dropping the measurability requirements, but we do not pursue this.

If m is fixed, say $m = k$ and T_n does not depend on F then

$$\begin{aligned} L^*_{k,n}(x) = P^*\left(T^*_k \le x\right) &= \frac{1}{n^k}\sum_{i_1=1}^{n}\cdots\sum_{i_k=1}^{n} 1\left(T_k\left(X_{i_1},\ldots,X_{i_k}\right) \le x\right) \\ &\equiv V_k, \end{aligned}$$

where V_k is a V-statistic whose kernel has finite moments for all orders. Therefore V_k and its corresponding U-statistic have the same limiting distribution which is the expected value of the kernel Serfling [1980]. In general it is reasonable to expect that $T_k\left(X_1, \ldots X_k; \hat{F}_n\right)$ will behave like $T_k(X_1, \ldots X_k; F)$ where F can now be treated as fixed. In the theorem of this section we build this into the assumptions which needs to be verified for our major application.

For $m \to \infty$ and $m/n \to \lambda$ $(0 \le \lambda \le 1)$, set

$$\begin{aligned} U_n(\lambda) &= L^*_{[n\lambda]+1,n}, \quad 0 \le \lambda \le 1 - \frac{1}{n}, \\ U_n(\lambda) &= U_n\left(1 - \frac{1}{n}\right), \quad 1 - \frac{1}{n} < \lambda \le 1. \end{aligned} \tag{4}$$

Then, $U_n(\cdot)$ can be viewed as a stochastic process on $[0,1]$ whose values are $D(\bar{R})$ valued random elements on Ω. Equivalently, $U_n : [0,1] \times \Omega \to D(\bar{R})$.

With these notations, we consider the sequence of m's defined in (1) and assume:

(A.0) If $m = k$ fixed then $L^*_{k,n} \stackrel{\mathcal{L}}{\Rightarrow} L_k$ as $n \to \infty$, where

$$L_k(x) = P(T(X_1, \ldots, X_k, F) \leq x).$$

(A.1) $L_n \stackrel{\mathcal{L}}{\Rightarrow} L$ as $n \to \infty$, where L is a continuous distribution function. Viewed as random distribution function, L is fixed with probability 1 and belongs to $C(\bar{R})$, the continuous functions on $\bar{R}$ endowed with the sup norm, where $\bar{R} = [-\infty, +\infty]$.

(A.2) For $m \to \infty$, $m/n \to 0$, $L^*_{m,n} \stackrel{\mathcal{L}}{\Rightarrow} L$ in probability (see Bickel et al. [1997] for conditions).

(A.3) The L_k's defined in (A.0) are different for different fixed values of k.

(A.4) For all $(\lambda_1, \ldots, \lambda_l)$ $l \geq 1$,

$$(U_n(\lambda_1), \ldots, U_n(\lambda_l)) \stackrel{\mathcal{L}}{\Rightarrow} (U(\lambda_1), \ldots, U(\lambda_l))$$

in probability, where $U : [0,1) \times \Omega \to C(\bar{R})$. That is, if ρ is the product Skorokhod metric on $D(\bar{R}) \times \cdots \times D(\bar{R})$ then for suitable $\tilde{U}_n, U$ as in (a),(a') above

$$\rho\left(\left(\tilde{U}_n(\lambda_1), \ldots, \tilde{U}_n(\lambda_l)\right), (U(\lambda_1), \ldots, U(\lambda_l))\right) \xrightarrow{p} 0.$$

(A.5) $P\left(\lambda \mapsto U(\lambda) \text{ is } 1-1\right) = 1$.

With these preparations we can state and prove our first result.

Theorem 1 : Let $\hat{m}$ be the m chosen by the rule presented in Section 2. Then under assumptions (A.0)–(A.5):

$$\hat{m} \xrightarrow{p} \infty, \qquad \frac{\hat{m}}{n} \xrightarrow{p} 0.$$

Proof . For fixed $S < \infty$ by (A.0) and (A.3),

$$\min_{1\le s\neq k\le S} \left|\left|L^*_{s,n} - L^*_{k,n}\right|\right|_\infty \xrightarrow{p} \min_{1\le s\neq k\le S} ||L_s - L_k||_\infty > 0.$$

On the other hand, by (A.2), if $\langle\sqrt{n}\rangle$ is the m_j closest to $\sqrt{n}$ and similarly define $\langle q\sqrt{n}\rangle$ then,

$$\left|\left|L^*_{\langle\sqrt{n}\rangle,n} - L^*_{\langle q\sqrt{n}\rangle,n}\right|\right|_\infty \xrightarrow{p} 0. \tag{5}$$

Therefore,

$$\hat{m} \xrightarrow{p} \infty. \tag{6}$$

If $\tilde{L}^*_{m,n}$ correspond to $\tilde{U}_n(\cdot)$ just as $L^*_{m,n}$ correspond to $U_n(\cdot)$, define $\tilde{\hat{j}}$ in relation to $\{\tilde{L}^*_{m,n}\}$ as $\hat{j}$ is defined for $\{L^*_{m,n}\}$. Then, by construction of $\hat{m}$, for each J,

$$P\left(\tilde{\hat{j}} > J\right) \ge P\left(\min_{1\le j\le J} \left|\left|\tilde{L}^*_{m_j,n} - \tilde{L}^*_{m_{j+1},n}\right|\right|_\infty > \left|\left|\tilde{L}^*_{\langle\sqrt{n}\rangle,n} - \tilde{L}^*_{\langle q\sqrt{n}\rangle,n}\right|\right|_\infty\right) \tag{7}$$

for n sufficiently large.

If condition (a) for convergence in law of U_n holds in (A.4), then

$$\min_{1\le j\le J} \left|\left|\tilde{L}^*_{m_j,n} - \tilde{L}^*_{m_{j+1},n}\right|\right|_\infty \overset{\mathcal{L}}{\Rightarrow} \min_{1\le j\le J} \left|\left|U(q^j) - U(q^{j+1})\right|\right|_\infty. \tag{8}$$

Finally by (5), (7) and (8),

$$\liminf_n P\left(\tilde{\hat{j}} > J\right) \ge P\left(\min_{1\le j\le J} \left|\left|\, U\left(q^j\right) - U\left(q^{j+1}\right)\,\right|\right|_\infty > 0\right).$$

Therefore, for each J, $P\left(\tilde{\hat{j}} > J\right) \to 1$. But, if (a) holds for (A.4), by construction, $\tilde{\hat{j}}$ and $\hat{j}$ have the same distribution. Hence,

$$\frac{\hat{m}}{n} \xrightarrow{p} 0. \tag{9}$$

The result follows from (6) and (9). Similarly if (a') holds for (A.4) $\left|\left|\tilde{U}_n(\cdot) - U_n(\cdot)\right|\right|_\infty = o_p(1)$. Then, $P\left(\tilde{\hat{j}} \neq \hat{j}\right) \to 0$, and (9) follows for this case as well. □

3.2 Choice of m when the n-bootstrap works

This section is motivated by examples like the t statistic and $\sqrt{n}\left(\bar{X}-\mu\right)$, for which the n-bootstrap works Bickel and Freedman [1981],Singh [1981]. In such cases using bootstrap samples of size smaller than n, in general, causes a loss in efficiency Bickel et al. [1997] (a counter example is given in Sakov and Bickel [2000]). In this section, we consider the conditions under which the rule picks m, such that there is no loss in efficiency. This involves assumptions of higher order. Suppose that,

$$L_n(\cdot) = A_0(\cdot, F) + n^{-\alpha} A_1(\cdot, F) + o(n^{-\alpha}) \tag{10}$$

holds with $\alpha = 0.5$. Assume further

$$L^*_{m,n}(\cdot) = A_0\left(\cdot, \hat{F}_n\right) + m^{-\alpha} A_1\left(\cdot, \hat{F}_n\right) + o_p\left(m^{-\alpha}\right), \tag{11}$$

where

$$\left|\left|A_0\left(\hat{F}_n\right) - A_0\left(F\right)\right|\right| = \Omega_p\left(n^{-1/2}\right) \tag{12}$$

$$\left|\left|A_1\left(\hat{F}_n\right) - A_1\left(F\right)\right|\right| = \Omega_p(1), \tag{13}$$

and $A_1(F) \neq 0$. Recall that L_n is the law of $T_n(X_1, \ldots, X_n, F)$ under $F \in \mathcal{F}$. Let $\|\cdot\|$ be the norm on $\mathcal{F}$, viewed as a subset of a suitable Banach space, and used in our rule. Define,

$$\lambda^*_n(F) = \inf_j \left|\left|L^*_{m_j,n} - L_n\right|\right| \qquad \hat{\lambda}^*_n(F) = \left|\left|L^*_{\hat{m},n} - L_n\right|\right|. \tag{14}$$

Theorem 2 Assume that assumptions (A.0)–(A.3) of Theorem 1 are fulfilled, and that (10)–(13) hold with $\alpha = 1/2$.

(a) Then

$$\hat{\lambda}^*_n(F) = \Omega_p(n^{-1/2}) \qquad \text{and} \qquad \lambda^*_n(F) = \Omega_p(n^{-1/2}), \tag{15}$$

and as a consequence

$$\frac{\hat{\lambda}^*_n(F)}{\lambda^*_n(F)} = \Omega_p(1) \qquad \text{for all } F \in \mathcal{F},$$

(b) If we assume further that $A_0(\cdot, F)$ does not depend on F,

$$\left|\left|A_1(\hat{F}_n) - A_1(F)\right|\right| = \Omega_p(n^{-1/2}),$$

and in addition $o_p\left(n^{-1/2}\right)$ is actually $O_p\left(n^{-1}\right)$ then

$$\frac{\lambda_n^*(F)}{\hat{\lambda}_n^*(F)} \xrightarrow{p} 1 \qquad \text{and} \qquad \hat{\lambda}_n^*(F) = O_p\left(n^{-1}\right). \tag{16}$$

In the first case, which corresponds to statistics like $\sqrt{n}\left(\bar{X} - \mu\right)$, the m-bootstrap with m chosen by the rule behaves equivalently to the n-bootstrap, and both commit errors of order $n^{-1/2}$. In the second case, which corresponds to statistics like the t statistic, the n-bootstrap produces coverage probabilities which differ by $O\left(n^{-1}\right)$ from the nominal ones and our conclusion (for $\|\cdot\|_\infty$) is that the $\hat{m}$ out of n bootstrap achieves the same performance.

Proof . (a) From assumptions (A.0)–(A.3) it follows that under F, $\hat{m} \xrightarrow{p} \infty$ (Theorem 1). Under the additional assumptions:

$$\begin{aligned} \left|\left|L_{n,n}^* - L_n\right|\right| &= \left|\left|A_0\left(\hat{F}_n\right) - A_0(F) + n^{-1/2}\left(A_1\left(\hat{F}_n\right) - A_1(F)\right)\right.\right. \\ &\quad \left.\left. + o_p\left(n^{-1/2}\right)\right|\right| = \Omega_p(n^{-1/2}). \end{aligned}$$

On the other hand,

$$\left|\left|L_{m_j,n}^* - L_{m_{j+1},n}^*\right|\right| = m_j^{-1/2}\left|\left|A_1\left(\hat{F}_n\right)\left(1 - q^{-1/2}\right) + o_p(1)\right|\right|,$$

which is minimized by $\hat{m} = n(1 + o_p(1))$. Thus (15) follows.

(b) If $A_0(\cdot, F) = A_0(\cdot)$ and

$$\left|\left|L_{n,n}^* - L_n\right|\right| = n^{-1/2}\left|\left|A_1\left(\hat{F}_n\right) - A_1(F) + o_p(1)\right|\right| = O_p\left(n^{-1}\right)$$

while

$$\left|\left|L_{m_j,n}^* - L_{m_{j+1},n}^*\right|\right| = m_j^{-1/2}\left|\left|A_1\left(\hat{F}_n\right)\left(1 - q^{-1/2}\right) + o_p(1)\right|\right|,$$

which is minimized to first order by $\hat{m} = n(1 + o_p(1))$ and to second order by $\hat{m} = n + o(1)$. Then

$$\begin{aligned} \left|\left|L_{\hat{m},n}^* - L_n\right|\right| &= \left|\left|\hat{m}^{-1/2}\left(A_1\left(\hat{F}_n\right) - A_1(F)\right)\right.\right. \\ &\quad \left.\left. + \left(\hat{m}^{-1/2} - n^{-1/2}\right)A_1(F) + o_p\left(\hat{m}^{-1/2}\right)\right|\right| \\ &= O_p\left(n^{-1}\right) \end{aligned}$$

and (16) follows. □

Corollary 1 Under the conditions of Theorem 2, $\hat{m}/n \xrightarrow{p} 1$.

3.3 Optimality of $\hat{m}$ when the n-bootstrap is inconsistent

When the n-bootstrap is inconsistent, natural questions to ask are whether,

(1) $\hat{m}$ achieves the optimal rate for the m out of n bootstrap.

(2) If so, does $\hat{m}$ also achieve the minimax rate for estimating L_n for a suitable norm ?

Using the definitions in (14) question (1) becomes: Does

$$\frac{\hat{\lambda}_n^*(F)}{\lambda_n^*(F)} = \Omega_p(1) \qquad \text{for all } F \in \mathcal{F}.$$

Question (2) could be framed as follows. Assume

$$\rho_n = \inf_{\delta} \max_{\mathcal{F}} E_F \|\delta(X_1, \ldots, X_n) - L_n\| \to 0,$$

where δ ranges over all possible estimates of L_n . Then, does

$$\frac{\mathrm{E}_F \, ||L_{\hat{m}}^* - L_n||}{\rho_n} = \frac{\mathrm{E}_F \hat{\lambda}_n^*(F)}{\rho_n} = \Omega(1).$$

Götze [1993] address Question (1) for a class of smooth functionals $T(\hat{F}_n)$ and certain norms obtaining an affirmative answer.

We briefly address Question (1) here. The two questions are being also addressed in Section 4.2 in the context of extrema.

Our basic requirement for addressing Question (1) is the existence of Edgeworth type expansion for L_n and $L_{m,n}^*$. Assume (10) holds and

$$A_0(\cdot, \hat{F}_n) = A_0(\cdot, F) + O_p(n^{-\gamma}), \qquad A_1(\cdot, \hat{F}_n) = A_1(\cdot, F) + \Omega_p(1).$$

The expansion for $L_{m,n}^*$ is assumed to have a different form and is as follows:

$$L_{m,n}^*(\cdot) = A_0\left(\cdot, \hat{F}_n\right) + m^{-\alpha} A_1\left(\cdot, \hat{F}_n\right) + \Omega_p(m^{\beta} n^{-\gamma}) + o_p(m^{-\alpha} + m^{\beta} n^{-\gamma}) \tag{17}$$

Here $\Omega_p(m^{\beta} n^{-\gamma})$ is to be interpreted as valid for any sequence $m_n \to \infty$. When, for example, $\mu = 0$ the Edgeworth expansion of $\sqrt{n}\bar{X}_n$ has

$\alpha = 0.5$. Bootstrapping $\sqrt{m}\bar{X}^*_m$ has an expansion with $\beta = \gamma = 0.5$. This is also the case for the trimmed-mean Sakov [1998]. Contrasting (17) and (11) shows that the former has an additional term. In the case of the mean it would be $\sqrt{m}\bar{X}_n$. When $m = o(n)$ this term is negligible, but not when $m = n$. These terms reflect what we already know. For fixed m the bootstrap converges at the usual types of rates, $n^{-0.5}, n^{-1}$ etc. But, as $m \to \infty$, since the n bootstrap fails, there have to be m terms that blow up. How terms of the type we consider appear may be seen for instance in Sakov and Bickel [2000].

Theorem 3 1. Under (10) and (17),

$$\inf_m \left|\left|L^*_{m,n} - L_n\right|\right| = \Omega_p\left(n^{-\alpha\gamma(\alpha+\beta)^{-1}}\right). \tag{18}$$

2. If $\hat{m}$ is chosen by the rule then

$$\left|\left|L^*_{\hat{m},n} - L_n\right|\right| = \Omega_p\left(n^{-\alpha\gamma(\alpha+\beta)^{-1}}\right). \tag{19}$$

3. Let m_{opt} be the m which minimizes (18) then $m_{opt}/\hat{m} = \Omega_p(1)$.

Proof . Writing

$$\begin{aligned} L_n - L^*_{m,n} &= A_0(\cdot, F) - A_0(\cdot, \hat{F}_n) + A_1(\cdot, F)(n^{-\alpha} - m^{-\alpha}) \\ &\quad + o_p(m^{-\alpha}) + \Omega_p(m^{\beta}n^{-\gamma}) \\ &= A_1(\cdot, F)m^{-\alpha} + \Omega_p(m^{\beta}n^{-\gamma}) + o_p(m^{-\alpha}) + O_p(n^{-\gamma}), \end{aligned}$$

yielding an (optimal) minimizing value of $m_{opt} = \Omega_p(n^{\gamma(\alpha+\beta)^{-1}})$ and (18) follows. Similarly,

$$L^*_{m_{j+1},n} - L^*_{m_j,n} = A_1(F)m_j^{-\alpha}\left(\frac{1}{q^{\alpha}} - 1\right) + \Omega_p\left(m_j^{\beta}n^{-\gamma}\right) + o_p\left(m^{-\alpha}\right),$$

yielding $\hat{m} = \Omega_p\left(n^{\gamma(\alpha+\beta)^{-1}}\right)$ and (19) follows. The last claim follows from the above. □

4 Confidence bounds for extrema

We discuss the above conditions in the context of setting confidence bounds for extrema, which is our main example. Note that the assumptions of Götze and Račkauskas [2001] do not hold in this example.

Assume $X_1, \ldots, X_n$ is an iid sample from a distribution F, with density f, which is in the domain of attraction of G for the maximum. Then, G is of one the three types of extremal distributions for the maximum David [1981]. This means that if $X_{(n)} = \max(X_1, \ldots, X_n)$ then there are normalizing constants $a_n > 0$ and $b_n \in R$ (depending on F) such that

$$P\left(a_n(X_{(n)} - b_n) \leq x\right) = F^n\left(\frac{x}{a_n} + b_n\right) \to G(x), \tag{20}$$

for all x in the support of G. The three types are specified by

$$\begin{aligned} &\text{(I)} && G(x) = e^{-x^{-\gamma}},\ x \geq 0,\ \gamma > 0. \\ &\text{(II)} && G(x) = e^{-(-x)^{\gamma}},\ x \leq 0,\ \gamma > 0. \\ &\text{(III)} && G(x) = e^{-e^{-x}}. \end{aligned} \tag{21}$$

From (5.1.8)–(5.1.11) and P.5.3 in Reiss [1989] it follows that the constants a_n and b_n can be chosen to be,

$$b_n(F) = F^{-1}\left(1 - \frac{1}{n}\right),$$

and

$$a_n^{-1}(F) = \begin{cases} b_n, & \text{for type I} \\ \omega(F) - b_n, & \text{for type II} \\ \frac{1}{nf(b_n)}, & \text{for type III} \end{cases},$$

where $\omega(F)$ is the supremum of the support of F. With this choice of normalizing constants the limiting distribution in (20) might change to $G((x-\mu)/\sigma)$ for some μ and σ, i.e. a location-scale change, but the type remains the same.

The von Mises conditions gives the conditions on F to belong to the domain of attraction of G. We consider a slightly specialized form

of these conditions, see p. 159 of Reiss [1989]: Let $F'' = f'$ exist. Then, for $\gamma > 0$,

$$\text{If} \quad \lim_{x \uparrow \omega(F)} [(1-F)/f]'(x) = \begin{cases} \frac{1}{\gamma}, & \text{then converges to type I} \\ -\frac{1}{\gamma}, & \text{then converges to type II} \\ 0, & \text{then converges to type III} \end{cases} . \tag{22}$$

We shall refer to these conditions as (vM)(1)–(3) respectively. The original von Mises conditions, which are more tedious to work with and hard to check directly, are implied by (22). Our results can be proved under the original conditions as a referee has pointed out but since all common example satisfy the modified conditions we choose WAS to state all our results under (22).

Our goal now is to set an upper confidence bound on the unknown $b_n(F)$. If G and a_n are assumed to be known then an upper approximate confidence bounds on $b_n(F)$ based on $X_{(n)}$ is $X_{(n)} - a_n^{-1} G^{-1}(\alpha)$ (although the convergence to G can be slow). One natural alternative is to bootstrap the distribution of $a_n \left(X_{(n)} - F^{-1}\left(1 - 1/n\right) \right)$, and replace $G^{-1}(\alpha)$ by the α-th quantile of the bootstrap distribution of $a_n(X^*_{(n)} - \hat{F}_n^{-1}(1 - 1/n))$. Unfortunately, as is well known, the n-bootstrap, in this case, does not work Athreya and Fukuchi [1993]. Below we denote by $X_{(k,n)}$ the k-th order statistic of a sample of size n and by $X^*_{(k,m)}$ the k-th order statistic of a bootstrap sample of size m. If we apply our paradigm to

$$T_n = a_n \left(X_{(n,n)} - F^{-1}\left(1 - \frac{1}{n}\right) \right), \tag{23}$$

we are led to consider the bootstrap distribution of

$$T^*_m = a_m \left(X^*_{(m,m)} - \hat{F}_n^{-1}\left(1 - \frac{1}{m}\right) \right) = a_m \left(X^*_{(m,m)} - X_{([n - \frac{n}{m}],n)} \right). \tag{24}$$

Denote the α-th quantile of this distribution by $\left(\hat{G}^*_m\right)^{-1}(\alpha)$, and use

$$X_{(n)} - a_n^{-1} \left(\hat{G}^*_m\right)^{-1}(\alpha) \tag{25}$$

as an upper confidence bound for $b_n(F)$ where $m \to \infty$, $m/n \to 0$.

We, naturally, propose to use $\hat{m}$ for m. The success of this strategy in the sense of Theorem 1 is given in the following section.

4.1 Order of $\hat{m}$

Theorem 4 Suppose F satisfies the von Mises conditions given above. Suppose also that the search for $\hat{m}$ in the rule is restricted to j such that $m_j = [q^j n]$ and $\left[\frac{n}{m_j}\right]$ are distinct integers as j varies. Then,

$$\mathcal{L}^* \left(T^*_{\hat{m}}\right) = \mathcal{L}^* \left(a_{\hat{m}} \left(X^*_{(\hat{m},\hat{m})} - X_{\left([n-\frac{n}{\hat{m}}],n\right)}\right)\right) \to G$$

in law in probability where G is the limit law of T_n defined in (23).

Proof . In order to prove the theorem, we show below that assumptions (A.0)–(A.5) of Section 3.1 hold. The theorem then follows from Theorem 1.

Condition (A.0): is immediate by continuity of $x \to T^*_m$.

Condition (A.1): is the classic result of von Mises Reiss [1989].

Condition (A.2): This condition is established if for $m \to \infty$ and $m/n \to 0$, the bootstrap statistic, T^*_m, is shown to weakly converge in probability to the same limiting distribution as T_n. Write,

$$T^*_m = a_m \left(X^*_{(m,m)} - b_m\right) - a_m \left(X_{\left([n-\frac{n}{m}],n\right)} - b_m\right).$$

It was shown in Athreya and Fukuchi [1993] that $a_m \left(X^*_{(m,m)} - b_m\right)$ converges weakly to the desired limit. In Lemma 1 (See the Appendix) we show that the right hand side of the last equation is $o_p(1)$, and condition (A.2) follows.

Condition (A.3): Bickel and Sakov [2002a] showed that if F satisfies (vM)(1)–(3) then (A.3) is valid unless F is an extreme value distribution, in which case $L_1(F) = L_2(F) = \cdots$. The condition follows from this result. In the exceptional case, that F is itself one of the three types, we only need to show $\hat{m}/n \xrightarrow{p} 0$ which requires conditions (A.4) and (A.5) only.

Condition (A.4): The proof of this condition uses the Poisson approximation to the Binomial distribution. A few notations and def-

initions are needed for that purpose. Let

$$\hat{p}_n(x,\lambda) = 1 - \hat{F}_n\left(\frac{x}{a_m} + b_m\right),$$

and

$$V_n(\lambda, x) = N(m\hat{p}_n(x,\lambda)), \tag{26}$$

where N denote a standard Poisson process independent of $X_1, X_2, \ldots$. Hence,

$$P^*\left(V_n(\lambda,x) = 0\right) = e^{-m\hat{p}_n(x,\lambda)}.$$

Denote, further,

$$V_n^*(\lambda,x) = \begin{cases} \sum_{i=1}^m 1\left(X_i^* > \frac{x}{a_m} + b_m\right), & m = [n\lambda]+1,\ 0 \le \lambda < 1 - \frac{1}{n} \\ V_n^*\left(1-\frac{1}{n}, x\right), & 1-\frac{1}{n} \le \lambda \le 1 \end{cases}, \tag{27}$$

Define $W_n(\lambda)$ to be the process

$$W_n(\lambda) = a_m\left(X_{\left([n-\frac{n}{m}],n\right)} - b_m\right).$$

Note that, for $m = \lambda n$,

$$\begin{aligned} U_n(\lambda,x) &= P^*(T_m^* \le x) = P^*\left(a_m\left(X^*_{(m,m)} - X_{\left([n-\frac{n}{m}],n\right)}\right) \le x\right) \\ &= P^*\left(V_n^*(\lambda, x + W_n(\lambda)) = 0\right). \end{aligned} \tag{28}$$

We need to define two more processes:

$$S_n(x,\lambda) = n\hat{p}_n(x,\lambda) = \sum_{i=1}^n 1\left(X_i > \frac{x}{a_m} + b_m\right),$$

and

$$\tilde{S}_n(x,\lambda) = S_n(x + W_n(\lambda), \lambda).$$

We set $\underline{S_n}(\cdot) = (S_n(\cdot,\lambda_1), \cdots, S_n(\cdot,\lambda_k))$ and similarly for $\underline{\tilde{S}_n}(\cdot)$ and $\underline{W_n}$. In Lemma 2 (See Appendix) we show that all $\lambda_1, \ldots, \lambda_k$, $k < \infty$

$$\rho(\mathcal{L}^*(V_n^*(\lambda_1), \ldots, V_n^*(\lambda_k)), \mathcal{L}^*(V_n(\lambda_1), \ldots, V_n(\lambda_k))) \xrightarrow{p} 0$$

where $\mathcal{L}^*$ is the joint conditional law and ρ is the Prohorov metric on probabilities on $D(\bar{R})$. From this result and the above definitions of

processes, it follows that

$$\rho \quad \left((U_n(\lambda_1), \ldots, U_n(\lambda_k)), (e^{-\lambda_1 n \hat{p}_n(\cdot + W_n(\lambda_1), \lambda_1)}, \ldots, e^{-\lambda_k n \hat{p}_n(\cdot + W_n(\lambda_k), \lambda_k)}) \right)$$
$$\rho \quad \left((U_n(\lambda_1), \ldots, U_n(\lambda_k)), (e^{-\lambda_1 \tilde{S}_n(\cdot, \lambda_1)}, \ldots, e^{-\lambda_k \tilde{S}_n(\cdot, \lambda_k)}) \right) = o_p(1). \qquad (29)$$

Lemma 3 (See Appendix) shows that the process $\left(\underline{S_n}(\cdot), \underline{W_n}\right)$ converges weakly to a limit

$$(\underline{S}(\cdot), \underline{W}) \equiv (S(\cdot, \lambda_1), \ldots, S(\cdot, \lambda_k), W(\lambda_1), \ldots, W(\lambda_k)).$$

Hence, $\underline{\tilde{S}_n}(\cdot)$ converges weakly to $\underline{\tilde{S}}(\cdot) \equiv \left(\tilde{S}(\cdot, \lambda_1), \ldots, \tilde{S}(\cdot, \lambda_k) \right)$, where $\tilde{S}(x, \lambda) \equiv S(x + W(\lambda), \lambda)$ for $1 \leq j \leq k$.

Combining (29) and the first part of Lemma 3 shows that (A.4) is satisfied.

Condition (A.5): Follows from Lemma 5 (See appendix). □

4.2 Optimality of $\hat{m}$

Here is a class of situations where (10) and (17) hold for $a_n \left(X_{(n)} - b_n \right)$. Let F satisfies the conditions of Theorem 5.2.11 (p. 176) of Reiss [1989]. Then (10) holds with α as given there for the Hellinger norm and, hence, also the smaller L_∞ norm. To see that (17) holds, consider m which grows polynomially i.e. $m = \Omega(n^r)$ for some $0 < r < 1$, then note that

$$\begin{aligned} \left|\left| L^*_{m,n} - L_m \right|\right| &= \sup_x \left\{ \left| \hat{F}_n^m \left(\frac{x}{a_m} + b_m \right) - F^m \left(\frac{x}{a_m} + b_m \right) \right| : x \in C(\epsilon) \right\} \\ &\quad + o\left(mn^{-1/2} \right) = \Omega_p(mn^{-1/2}). \end{aligned} \qquad (30)$$

where $C(\epsilon) = \left\{ x : F\left(\frac{x}{a_m} + b_m \right) \geq \epsilon \right\}$ for ϵ arbitrarily small. The argument for (30) is as follows. First,

$$\begin{aligned} &\sup_x \left\{ m \cdot \left| \log \hat{F}_n \left(\frac{x}{a_m} + b_m \right) - \log F \left(\frac{x}{a_m} + b_m \right) \right| : x \in C(\epsilon) \right\} \\ &= \Omega_p \left(m \sup_x \left\{ \left| \hat{F}_n(x) - F(x) \right| : x \in C(\epsilon) \right\} \right) = \Omega_p(mn^{-1/2}), \end{aligned}$$

and

$$\begin{aligned}\sup\left\{\left|\hat{F}_n^m\left(\frac{x}{a_m}+b_m\right)-F^m(\frac{x}{a_m}+b_m)\right|:x\notin C(\epsilon)\right\} &= o_p(\epsilon^m)\\ &= o_p(mn^{-1/2}).\end{aligned}$$

Using (30) and (10) we obtain that

$$\begin{aligned}L^*_{m,n}-L_n &= (L^*_{m,n}-L_m)+(L_m-L_n)\\ &= A_1(\cdot,F)\left(\frac{1}{m^\alpha}-\frac{1}{n^\alpha}\right)+o_p\left(m^{-\alpha}\right)+\Omega_p\left(mn^{-1/2}\right)\\ &= O_p\left(m^{-\alpha}\right)+\Omega_p\left(mn^{-1/2}\right),\end{aligned}$$

so that (17) holds with $\beta=1$, $\gamma=\frac{1}{2}$ and α as above for $||\cdot||=||\cdot||_\infty$.

The answer to question (2) of Section 3.3 "Are the rates we obtain minimax optimal ?" is more complicated. Here is what follows in a special case. Consider the case where we are dealing with a type II limit with $\omega(F)<\infty$ and unknown. Suppose that F has support on $[0,1]$ and F is such that $0<f(1-)<\infty$ and $|f'(1-)|\leq M<\infty$. Then it is well known that the minimax rate for estimating $f(1)$ using root mean square error is $n^{-1/3}$ but since $\beta=1,\gamma=1/2$ and $\alpha=1$, the m out of n bootstrap with $\hat{m}$ chosen as above can only yield a rate of $n^{-1/4}$.

This is a consequence of our using $||\cdot||_\infty$ as our measure of departure of $L^*_{m,n}$ from L_n, and of $L^*_{m_j+i,n}$ from $L^*_{m_j,n}$. If instead, we use the norm $||g||\equiv\sup\{|g(x)|:|x|\leq M\}$ for $M<\infty$, it may be shown that we obtain a rate of $n^{-1/3}$ for $||L_{\tilde{m},n}-L_n||$, where $\tilde{m}$ is selected using the same norm. This implies that a better estimate of L_n even in the original $||\cdot||_\infty$ is to estimate $f(1-)$ using $L_{\tilde{m},n}$ and then plug the resulting estimate into the limiting exponential distribution.

From a statistical point of view, there is an unfortunate general conclusion. We can design a pivot

$$a_n\left(\hat{F}_n\right)\left(X_{(n)}-F^{-1}\left(1-\frac{1}{n}\right)\right)$$

which has a limiting distribution independent of F in the domain of attraction of a particular type of extremal law. However, unlike what

happens with pivots such as the t statistic, the $\hat{m}$ out of n bootstrap distribution of this pivot is not theoretically a better approximation to the law of the pivot than the limit is.

4.3 The Case a_n Unknown

Throughout the paper we assume that the rate of convergence in known. Consider the case when the rate is unknown but has the following form: $a_n/n^\alpha \to c$, for some $\alpha \geq 0$, $c > 0$. Bertail et al. [1999] proposed estimating α using the m out of n bootstrap. They discuss this method for a general statistic, but we applied it to the current problem of extrema (as a preliminary stage, before applying the rule). In the context of extrema this form covers many important cases, e.g., F is the exponential, the uniform, generalized Pareto etc. However, it does rule out cases like F being Gaussian, where $a_n = \sqrt{2 \log n}$. Bertail et al. [2004] extends subsampling to a diverging statistic, and apply the rate estimation to an extrema problem.

5 Simulations

5.1 Choice of m

In this part of the paper we demonstrate, visually, that our choice of m makes sense and that moderate sample behavior reflects the asymptotics. The m chosen by the rule is the argmin of the *successive differences* $\left\| L^*_{m_j,n}(\cdot) - L^*_{m_{j+1},n}(\cdot) \right\|_\infty$. By assumption, $L_n \Rightarrow L$, so ideally we would like to search for the argmin of the *exact differences* $\left\| L^*_{m_j,n}(\cdot) - L(\cdot) \right\|_\infty$. In the simulation we have the privilege of knowing L (or estimating it using Monte-Carlo), so we can plot the successive and exact differences vs. m, which is what we do in Figure 1. We can then compare the m's which minimize the two curves.

As can be seen from the figures there are two "typical" types of curves for the successive and exact differences: when the n-bootstrap fails and when it works. Hence this provides also a diagnostic for n-bootstrap failure. When the bootstrap fails, the successive differences go down drastically, and then increase. On the other hand, when the n-bootstrap works, the successive differences drop down and then remain pretty constant. This reflects the fact that when the n-bootstrap works, once m is large enough there will not be differences, to first order, between bootstrap distributions at increasing sample sizes. We give four situations: three when the bootstrap fails and one when the bootstrap works.

1. The extreme as discussed in Section 4 when the data follows the exponential distribution.
2. The extreme when the data follows the Gumbel distribution but assuming a_n is unknown. a_n is being estimated using the approach of Bertail et al. [1999] (for more details see 2 in Section 5.2).
3. The classical example of bootstrap failure discussed in Bickel and Freedman [1981]: assume the data follows the the Uniform$(0, \theta)$ distribution. The mle of θ is $X_{(n)}$, the n-bootstrap fails to estimate its distribution. Using the m-bootstrap helps.

4. Bootstrapping the normalized mean (t-statistic) in which the n-bootstrap works Bickel and Freedman [1981],Singh [1981].

The sample size is indicated above each plot. The number of bootstrap samples is 1,000 and the plots were generated using $q = 0.75$. Note that the two curves are not always close to each other, however, both achieve their minimum for about the same m, and this is the purpose of the rule – to pick m.

For more plots we refer the reader to Sakov [1998], Sakov and Bickel [1999], Götze and Račkauskas [2001].

5.2 Coverage when m is chosen by the rule

In this part we compute the coverage when estimating a parameter, using the m-bootstrap with m chosen by the rule.

1. In table 1 an upper bound for $b_n = F^{-1}(1 - 1/n)$ was computed using the approach discussed in Section 4. The statistic is (23), and its bootstrap version is (24). Four distributions were considered: Exponential, Normal and Gumbel (domain of attraction is type III) and Uniform distribution (domain of attraction is type II). The a_n used are taken from Embrechts et al. [1997]. For all, but the exponential distribution, $F^{-1}(1 - 1/n)$ is different from the b_n given in Embrechts et al. [1997]. However, this does not change the domain of attraction, but potentially change the location parameter of the limiting distribution. The sample sizes are 500, 1000 and 10,000. The upper bound was estimated in 1000 repetitions, and the coverage is evaluated. The mean, SD and median of the chosen m over the 1000 repetitions are given as well.
2. Table 2 is similar, but this time the normalizing constant a_n is unknown, but is assumed to have the form $a_n = n^\alpha$ for some α. In each repetition α was estimated using the approach of Bertail et al. [1999]. Once the rate was estimated, a_n and a_m in (23) and (24) were replaced with $n^{\hat{\alpha}}$ and $m^{\hat{\alpha}}$ respectively. There are

two alternative methods to estimate α. We used the method which is based on ranges with $I = 15$, $J = 50$ and quantiles between $(0.75, 0.95)$ and $(0.05, 0.25)$. Note that a_n for the normal distribution does not have the desired form. Table 2 shows also the mean and SD of the α found over the 1000 repetitions.

3. In Table 3 the coverage was estimated in three situations.
   (a) An upper bound for μ when the variance is 1 is $\bar{X}_n - z_\alpha/\sqrt{n}$. The n-bootstrap works Singh [1981],Bickel and Freedman [1981]. m is chosen by the rule and the m-bootstrap replaces the normal quantile by the bootstrap quantile of $\sqrt{m}\left(\bar{X}^*_m - \bar{X}_n\right)$. The distributions considered were the Normal and Exponential (shifted to have a 0 mean).
   (b) An upper bound μ when the variance is unknown is $\bar{X}_n - t_{n-1,\alpha}\, s_n/\sqrt{n}$. Here, as well, the n-bootstrap works Singh [1981],Bickel and Freedman [1981], and the t-quantile is being replaced by the quantile of the bootstrap distribution of $\sqrt{m}\left(\bar{X}^*_m - \bar{X}_n\right)/s^*_m$.
   (c) An upper bound for θ when the data follows the Uniform$(0, \theta)$ is $X_{(n)} - \log(\alpha)/n$ (the bound is based on $n(\theta - X_{(n)})$). This is a classical example of n-bootstrap failure Bickel and Freedman [1981]. Using the bootstrap, the exponential quantile is being replaced by the quantile of the bootstrap distribution of $m(X_{(n)} - X^*_{(m)})$.

In all the simulations the coverage was evaluated by repeating the process 1,000 times and checking how many of the 1,000 bounds covered the known value of the parameter. The number of bootstrap samples was 1,000, and the sequence of m was generated using $q = 0.75$. The desired level was 95%.

5.3 Choice of q, choice of metric and smoothing

In the simulations in this paper we have used $q = 0.75$ in the sequence (1). In Sakov [1998] experiments with $q = 0.5$ gave qualitatively the

same answers. In the extrema problem, coverage for b_n was about the same for $q = 0.75$ and $q = 0.5$ (differences ranged between 0.004 and 0.04 and typically were around 0.01). This was the case for the 95% bound as well as for the 90% bound.

A referee suggested, on theoretical grounds, that a metric based on comparison of the appropriate quantiles of the bootstrap distributions may be preferable over the Kolmogorov sup distance. We noted, in Section 3.3, that a restricted Kolmogorov sup distance should also give better results for the upper confidence bound for the upper end of the support in the case of bounded Lipschitz densities.

In the simulation presented here we have concentrated on Kolmogorov sup distance. In addition, we studied the Wasserstein metric with $p = 1$ and 2 as well as a quantile-based metric. In simulation, the cdf is evaluated on a finite grid over a bounded interval so the Kolmogorov sup distance is actually a restricted Kolmogorov distance as is the Wasserstein metric. When using the Wasserstein metric with $p = 1$ and 2 coverage was about 1-2% smaller than the coverage achieved using the Kolmogorov sup distance for the exponential distribution. For the Normal distribution the coverage using the Wasserstein metric was about 3-4% lower. We also picked m which minimizes the distance between the 5% quantile of the bootstrap distributions. Here the coverage was about 1 - 1.5% lower than the coverage using the Kolmogorov sup distance (for the exponential, normal and uniform distributions). To verify the claim made in Section 3.3 we evaluated the cdf of $n(X_{(n)}-1)$ over the grids $(-5, 0)$ and $(-20, 0)$, expecting that the latter interval should behave like the unrestricted Kolmogorov sup distance. For sample sizes of 1000 and 4000 the coverage probabilities using the smaller interval were 0.2 - 1% better.

Following a suggestion of the referee we explored the effect of smoothing the curve whose minima we find before locating them. In practice, we smoothed the curve using local polynomials with span of 30% (i.e. if the sequence of m's includes K values then each window contained 30% of them), and then picked m which minimized the smoothed curve. We did this for the Kolmogorov sup distance and found that differences in

coverage were less than 1%.

All in all these variants did not make any substantial differences in performance for the sample sizes we considered.

Acknowledgment: The authors would like to thank the referee for a thorough report. The comments improved the manuscript significantly.

References

K. Athreya and J. Fukuchi. Bootstrapping extremes of i.i.d. random variables. In *Proceedings of the conference on extreme value theory and applications*, volume 3, pages 23–29, 1993.

A. Barbour. Stein's method and poisson process convergence. *Journal of Applied Probability*, Special volume 25A:175–184, 1988.

R. Beran. Estimated sampling distributions: the bootstrap and competitors. *The Annals of Statistics*, 10:212–225, 1982.

P. Bertail, C. Haefke, D.N. Politis, and H. White. Subsampling the ditrbiution of diverging statistics with applications to finance. *Journal of Econometrics*, 120:295–326, 2004.

P. Bertail, D. Politis, and J. Romano. On subsampling estimators with unknown rate of convergence. *Journal of the American Statistical Association*, 94:569–579, 1999.

P. Bickel and D. Freedman. Some asymptotic theory for the bootstrap. *The Annals of Statistics*, 9:1196–1217, 1981.

P. Bickel, F. Götze, and W. van Zwet. Resampling fewer than n observations: gains, losses and remedies for losses. *Statistica Sinica*, 7: 1–31, 1997.

P. Bickel and A. Sakov. Equality of types for the distribution of the maximum for two values of n implies extreme value type. *Extremes*, 5:45 – 53, 2002a.

P. Bickel and A. Sakov. Extrapolation and the bootstrap. *Sankhya*, 64 (Basu memorial volume):640–652, 2002b.

J. Bretagnolle. Lois limites du bootstrap de certaines fonctioneles. *Annales Inst. Henri Poincare*, 19:281–296, 1983.

S. Datta and W.P. McCormick. Bootstrap inference for a first-order autoregression with positive innovations. *Journal of the American statitical Association*, 90:1289–1300, 1995.

H. David. *Order statistics*. John Wiley & Sons, 2nd edition, 1981.

B. Efron. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7:1–26, 1979.

P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling extremal events*. Springer, 1997.

D. Freedman. Bootstrapping regression models. *The Annals of Statistics*, 9:1218–1228, 1981.

F. Götze. Asymptotic approximation and the bootstrap. IMS Bulletin, 1993. p.305.

F. Götze and A. Račkauskas. Adaptive choice of bootstrap sample sizes. In *State of the art in probability and statistics*, I.M.S lectures notes Ser. 36, pages 286–309. I.M.S publications, 2001.

P. Hall, J.L. Horowitz, and B. Jing. On blocking rules for the bootstrap with dependent data. *Biometrika*, 82:561–574, 1995.

J. Hodges and L. LeCam. The Poisson approximation to the Poisson binomial distribution. *The Annals of Mathematical Statistics*, 31: 737–740, 1960.

D. Politis and J. Romano. Large sample confidence regions on subsamples under minimal assumptions. *The Annals of Statistics*, 22: 2031–2050, 1994.

D. Politis, J. Romano, and M. Wolf. *Subsampling*. Springer, 1999.

R. Reiss. *Approximate distributions of order*. Springer, 1989.

A. Sakov. *Using the m out of n bootstrap in hypothesis testing*. PhD thesis, University of California, Berkeley, 1998.

A. Sakov and P. Bickel. Choosing m in the m out of n bootstrap. In *Proceedings of the American Statistical Association, Section on Bayesian Statistical Science*, pages 125–128, 1999.

A. Sakov and P. Bickel. An Edgeworth expansion for the m out of n bootstrapped median. *Statistics and Probability Letters*, 49:217–223, 2000.

R. Serfling. *Approximation theorems to Mathematical Statistics*. Wiley, 1980.

B. Silverman. Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Soc. B*, 43:97–99, 1981.

K. Singh. On the asymptotic accuracy of Efron's bootstrap. *The Annals of Statistics*, 9:1187–1195, 1981.

A. SintesBlanc. On the poisson approximation for some multinomial distributions. *Statistics and Probability letters*, 11:1–6, 1991.

J. Swanepoel. A note on proving that the (modified) bootstrap works. *Communication in Statistics, Part A*, 15:3193–3203, 1986.

A. van der Vaart and J. Wellner. *Weak convergence and empirical processes. With applications to Statistics*. Springer-Verlag, 1996.

Y. Wang. Coupling methods in statistics. *The Canadian journal of Statistics*, 14:69–74, 1986.

A Appendix

Lemma 1 : Under (vM)(1)–(3), if $m \to \infty$, $m/n \to 0$ then

$$a_m \left(X_{([n-\frac{n}{m}],n)} - b_m \right) = o_p(1). \tag{31}$$

Proof . From Theorem 5.1.7 (p. 164) of Reiss [1989], under (vM)(1)–(3) (and identifying $k(n)$ with $n - \left[n - \frac{n}{m}\right]$, and Reiss's b_n is our b_m) it follows that

$$(nm)^{\frac{1}{2}} f(b_m) \left(X_{([n-\frac{n}{m}],n)} - b_m \right) \tag{32}$$

converges weakly to the standard normal distribution. In case (vM)(3), $a_m = mf(b_m)$ and the lemma follows. In case (vM)(1) $a_m = \frac{1}{b_m}$ and from (5.1.24) of Reiss [1989] $mb_m f(b_m) = O(1)$, so

$$\frac{a_m}{\sqrt{nm} f(b_m)} = \frac{1}{b_m \sqrt{m} f(b_n) \sqrt{n}} = o(1),$$

and the lemma follows. In case (vM)(2), and using (5.1.24) of Reiss [1989], the result follows using a similar argument. □

Lemma 2 : If (20) holds then for all $\lambda_1, \ldots, \lambda_k$, $k < \infty$

$$\rho(\mathcal{L}^*(V_n^*(\lambda_1), \ldots, V_n^*(\lambda_k)), \mathcal{L}^*(V_n(\lambda_1), \ldots, V_n(\lambda_k))) \xrightarrow{p} 0$$

where ρ is the Prohorov metric on probabilities on $D(\bar{R})$.

Proof . Denote by $\|\cdot\|_{BV}$ the total variation norm between two probability measures. In the proof of Lemma 2 we use the following proposition which is an extension of a result of Hodges and LeCam [1960] (for a proof see Barbour [1988], Wang [1986], SintesBlanc [1991]).

Proposition 1 : Let $\mathbf{X}_i = (X_{i1}, \ldots, X_{ik})$, $i = 1, \ldots, n$ be independent and distributed according to the multinomial distribution with parameters $(1, q_1, \ldots, q_k)$ such that $q_j \geq 0$ and $\sum_{j=1}^k q_j < 1$. Let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{ik})$, $i = 1, \ldots, n$ where Y_{ij} are independent Poisson random variables with means q_j. Then

$$\|\mathcal{L}(\mathbf{X}_1, \ldots, \mathbf{X}_n) - \mathcal{L}(\mathbf{Y}_1, \ldots, \mathbf{Y}_n)\|_{BV} \leq n \left(\sum_{i=1}^k q_i \right)^2,$$

where $\mathcal{L}(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ is the joint law of $(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$.

Continue proof of Lemma 2. For all positive integers K, L define

$$A_{K,L} = \{(k,l) \mid 1 \le k \le K,\ 1 \le l \le L\}.$$

Since for each λ, $V_n^*(\lambda,\cdot)$, $V_n(\lambda,\cdot)$ are monotone decreasing in x, it is suffices to establish for all $\lambda_1 < \cdots < \lambda_I$, $x_J < \cdots < x_1$, I, J:

$$R \equiv \rho(\mathcal{L}^*(V_n^*(\lambda_i, x_j) : (i,j) \in A_{I,J}),\ \mathcal{L}^*(V_n(\lambda_i, x_j) : (i,j) \in A_{I,J})) \xrightarrow{p} 0.$$

Let m_i be the m associated with λ_i, and let

$$C_{ij} = \begin{cases} \left(\frac{x_1}{a_{m_i}} + b_{m_i}, \infty\right), & j = 1 \\ \left(\frac{x_j}{a_{m_i}} + b_{m_i}, \frac{x_{j-1}}{a_{m_i}} + b_{m_i}\right], & \text{j = 2,...,J} \end{cases}.$$

For a given i, the intervals C_{ij} are disjoint, but they are not necessarily so for different i's. We use the end-points of the C_{ij} when $1 \le i \le I$ and $1 \le j \le J$ to create a partition of the real line into disjoint intervals D_r where $1 \le r \le \beta$ and $1 \le \beta \le IJ$ (β is *not* an index but a parameter, whose size depends on whether there are IJ disjoint end-points or not). We denote the end-points of the D's by $D_r = (z_{r-1}, z_r]$ and $-\infty < z_1 \ldots < z_\beta < z_{\beta+1} = \infty$. Each z_r is associated with a pair (x_j, λ_i) for some i and j. Finally, let $\hat{p}_n(z_r)$ be $\hat{p}_n(x_j, \lambda_i)$ with the corresponding i and j. With this preparation, we are ready to rewrite,

$$\begin{aligned} V_n^*(\lambda_i, x_j) &= \sum_{k=1}^{m_i} 1\left(X_k^* > \frac{x_j}{a_{m_i}} + b_{m_i}\right) = \sum_{k=1}^{m_i}\sum_{l=1}^{j} 1(X_k^* \in C_{il}) \\ &= \sum_{k=1}^{n}\sum_{r=1}^{\beta} \delta_{ijk} 1(X_k^* \in D_r), \end{aligned}$$

where

$$\delta_{ijk} = \begin{cases} 1, & \text{if } 1 \le k \le m_i \text{ and } D_r \subset \cup_{l=1}^{j} C_{il} \\ 0, & \text{otherwise} \end{cases}.$$

Now,

$$P^*\left(X_k^* \in D_r\right) = \hat{p}_n(z_{r-1}) - \hat{p}_n(z_r) \quad r = 1, \ldots, \beta.$$

Note that we can similarly write

$$V_n(\lambda_i, x_j) = \sum_{k=1}^{n}\sum_{r=1}^{\beta} \delta_{ijk} P_{kr},$$

where the δ_{ijk} are as above, and the P_{kr} are independent Poisson with $E(P_{kr}) = P^*(X_k^* \in D_r)$. Therefore,

$$\begin{aligned}
&||\mathcal{L}^*\left(\{V_n^*(\lambda_i, x_j) : (i,j) \in A_{I,J}\}\right) - \mathcal{L}^*\left(\{V_n(\lambda_i, x_j) : (i,j) \in A_{I,J}\}\right)||_{BV} \\
&\leq ||\mathcal{L}^*\left(1(X_k^* \in D_r) : (k,r) \in A_{n,\beta}\right) - \mathcal{L}\left(P_{kr} : (k,r) \in A_{n,\beta}\right)||_{BV} \\
&\leq n\left(\sum_{r=1}^{\beta} P^*(X_1^* \in D_r)\right)^2 \leq nIJ \max_{i,j}\{\hat{p}_n^2(x_j, \lambda_i)\},
\end{aligned}$$

where we have used the bound from Proposition 1. Since $n\hat{p}_n$ has the Binomial distribution:

$$\begin{aligned}
nE\hat{p}_n(x_j, \lambda_i) &= n\left(1 - F\left(\frac{x_j}{a_{m_i}} + b_{m_i}\right)\right), \\
\mathrm{Var}(n\hat{p}_n(x_j, \lambda_i)) &\leq n\left(1 - F\left(\frac{x_j}{a_{m_i}} + b_{m_i}\right)\right).
\end{aligned}$$

Note that (20) is equivalent to

$$n\left(1 - F\left(\frac{x}{a_n} + b_n\right)\right) \to -\log G(x). \tag{33}$$

With the above and since $n/m_i = O(1)$, we conclude that

$$n \max_{i,j} \hat{p}_n(x_j, \lambda_i) = O_p(1).$$

Thus,

$$R \leq \|\mathcal{L}^*\{V_n^*(\lambda_i, x_j) : (i,j) \in A_{I,J}\}) - \mathcal{L}^*(\{V_n(\lambda_i, x_j) : (i,j) \in A_{I,J}\})\|_{BV} = o_p(1),$$

and the lemma follows. □

Lemma 3 : Let N' be a standard Poisson process independent of N and of $X_1, X_2, \ldots$. Denote $\psi(x) = -\log G(x)$.

(i) The process $\left(\underline{S_n}(\cdot), \underline{W_n}\right)$ converges weakly to a limit

$$(\underline{S}(\cdot), \underline{W}) \equiv (S(\cdot, \lambda_1), \ldots, S(\cdot, \lambda_k), W(\lambda_1), \ldots, W(\lambda_k)).$$

Hence, $\underline{\tilde{S}_n}(\cdot)$ converges weakly to $\underline{\tilde{S}}(\cdot) \equiv \left(\tilde{S}(\cdot, \lambda_1), \ldots, \tilde{S}(\cdot, \lambda_k)\right)$, where $\tilde{S}(x, \lambda) \equiv S(x + W(\lambda), \lambda)$ for $1 \leq j \leq k$.

(ii) For $\underline{u} = (u_1, \ldots, u_k)$, let $N'(\cdot|\underline{u})$ denote the conditional distribution of $N'(\cdot)$ given $\tau_{r_j} = \psi(\sigma_j u_j + \mu_j)$ for $1 \leq j \leq k$ where (a) $\tau_1 < \tau_2 < \cdots$ are the consecutive jumps of $N'(\cdot)$: that is $N'(\tau_r) = r$ and $N'(\tau_r -) = r - 1$. (b) $\sigma_\lambda = \sigma(\lambda)$ and $\mu_\lambda = \mu(\lambda)$ are defined by $\sigma_\lambda = 1, \mu_\lambda = \log(\lambda)$ for (vM)(3) or $\sigma_\lambda = \lambda^{-2}, \mu_\lambda = 1 - \lambda^{-2}$ for (vM)(1)-(2).
Then $\left\{\tilde{S}(\cdot, \lambda_j) : j = 1, \ldots, k\right\}$ is distributed as $\{N'(\psi(\cdot + \sigma_j u_j + \mu_j)) : 1 \leq j \leq k\}$ given $W(\lambda_j) = u_j$ for $1 \leq j \leq k$.

Agenda for the proof of Lemma 3:
To prove Lemma 3 we shall argue that (a) $\underline{W_n}$ has a weak limit $\underline{W}$. (b) The conditional distribution of $\underline{S_n}(\cdot)$ given $\underline{W_n} = \underline{u} = (u_1, \ldots, u_k)$ convergences weakly to a measure on $D^k[-\infty, \infty]$, $Q_{\underline{u}}$. Then, $\left(\underline{S_n}(\cdot), \underline{W_n}\right)$ necessarily converges weakly to $(\underline{S}(\cdot), \underline{W})$, where $\underline{S}(\cdot)$ has conditional distribution $Q_{\underline{W}}$. (c) With these identifications it then follows that $\underline{\tilde{S}_n}(\cdot)$ given $W_n(\lambda_j) = u_j$ for $1 \leq j \leq k$ converges weakly to $(S(\cdot + u_1, \lambda_1), \ldots, S(\cdot + u_k, \lambda_k))$. From this we obtain the joint distribution of $\underline{\tilde{S}}(\cdot)$, and identify it in a fashion useful for Lemma 5. We begin with an auxiliary lemma.

Lemma 4 Under the von Mises conditions, if $m/n \to \lambda > 0$ then:

1. For (vM)(1) and (vM)(2), $a_m/a_n \to \lambda^2$ and $a_n(b_m - b_n) \to \frac{1}{\lambda^2} - 1$.
2. For (vM)(3) $a_m/a_n \to 1$ and $a_n(b_m - b_n) \to -\log(\lambda)$.

Proof . To simplify the notation, we denote m/n by λ rather then $\lambda^{(n)} \to \lambda$. The result is unaffected.

We start with (vM)(3), then $a_m = mf(b_m)$ and $a_n = nf(b_n)$. Noting that $1 - F(b_m) = \frac{1}{\lambda n}$ and then using (22), we obtain that

$$\begin{aligned}\frac{\partial \log(f(b_m))}{\partial \lambda} &= \frac{\partial \log f}{\partial \lambda}\left(F^{-1}\left(1 - \frac{1}{\lambda n}\right)\right) = \frac{f'(b_m)}{\lambda m f^2(b_m)} \\ &= -\frac{1}{\lambda}\left[\left(\frac{1 - F(b_m)}{f(b_m)}\right)' + 1\right] \to -\frac{1}{\lambda}.\end{aligned}$$

Hence,

$$\begin{aligned}\log\left(\frac{a_n}{a_m}\right) &= \log\left(\frac{f(b_n)}{\lambda f(b_m)}\right) = -\int_\lambda^1 \frac{dz}{z} + o(1) - \log(\lambda) \\ &= -\log(1) + \log(\lambda) - \log(\lambda) + o(1) \to 0,\end{aligned}$$

It follows that, $\frac{a_n}{a_m} \to 1$. Similarly,

$$\begin{aligned}\frac{\partial}{\partial\lambda}a_n\left(b_m - b_n\right) &= \left[nf(b_n)\left(F^{-1}\left(1-\frac{1}{\lambda n}\right) - F^{-1}\left(1-\frac{1}{n}\right)\right)\right]' \\ &= \frac{1}{\lambda^2}\frac{f\left(b_n\right)}{f\left(b_m\right)} = \frac{1}{\lambda}\frac{a_n}{a_m} \to \frac{1}{\lambda}\end{aligned}$$

Consider now (vM)(1) then $a_n = 1/b_n$ and $a_m = 1/b_m$. From (22) it follows that

$$\frac{\partial \log(f(b_m))}{\partial\lambda} \to -\frac{1}{\lambda}\left(\frac{1}{\lambda}+1\right). \tag{34}$$

Using (5.1.24) of Reiss [1989], which is an alternative set of sufficient conditions, it follows that

$$\frac{b_m f(b_m)}{1-F(b_m)} \cdot \frac{1-F(b_n)}{b_n f(b_n)} = \frac{b_m}{b_n} \cdot \frac{f(b_m)}{f(b_n)} \cdot \frac{m}{n} \to 1.$$

Using (34) it follows that

$$\frac{a_n}{a_m} = \frac{b_m}{b_n} \to \frac{1}{\lambda^2},$$

hence, $a_m/a_n \to \lambda^2$ and $a_n(b_m - b_n) = b_m/b_n - 1 \to 1/\lambda^2 - 1$.

The argument for type (vM)(2) is very similar. □

Proof of Lemma 3. From (5.1.28) in Reiss [1989] it follows that for $m = \lambda^{(n)}n$ and $\lambda^{(n)} \to \lambda$,

$$\tilde{W}_n\left(\lambda^{(n)}\right) = a_n\left(X_{([n-n/m],n)} - b_n\right)$$

converges weakly to a limiting distribution which depends on λ and G and is given in (5.1.29) of Reiss [1989]. To simplify notation, from now on, we drop the superscript in $\lambda^{(n)}$. Note that,

$$W_n(\lambda) = \frac{a_m}{a_n}\tilde{W}_n(\lambda) + a_m(b_n - b_m).$$

In view of Lemma 4, $W_n(\lambda)$ converges weakly, say to $W(\lambda)$. To show that $\underline{\tilde{W}_n}$ and hence $\underline{W_n}$ converges weakly to some $\underline{W}$ is a slight extension of (5.1.28) in Reiss [1989]: to show this we only need to translate joint statements about extrema to joint statements about empirical distribution functions. This completes the proof of part (a) from the agenda.

To proceed with part (b) of the agenda we define,

$$N_n(u) \equiv S_n\left(\psi^{-1}(u), 1\right). \tag{35}$$

Note that $N_n(u)$ is a nondecreasing counting process and converges weakly to a standard Poisson process N', and that

$$S_n(x,\lambda) = N_n\left(\psi\left(x\frac{a_n}{a_m} + (b_m - b_n)a_n\right)\right). \tag{36}$$

Hence, $S(x,\lambda) = N'(\psi(x\sigma_j - \mu_j)$. Let $\tau_{1,n} < \cdots < \tau_{n,n}$ be the jump times of $N_n(\cdot)$. Evidently,

$$\tau_{j,n} = \psi\left(a_n\left(X_{(n-j+1,n)} - b_n\right)\right) \quad j = 1, \ldots, n.$$

Let,

$$r_j = \begin{cases} \left[\frac{1}{\lambda_j}\right] + 1, & \text{if } \frac{1}{\lambda_j} \text{ is not an integer} \\ \frac{1}{\lambda_j}, & \text{otherwise} \end{cases} \qquad \text{for} \quad j = 1, \ldots, k,$$

Then $\{r_j\}$ is a non-decreasing sequence of positive integers ($1 \le r_j \le n$) since $m_1 > \cdots > m_k$ and hence $\tau_{r_k,n} \ge \cdots \ge \tau_{r_1,n}$. Using the jump points rewrite:

$$W_n(\lambda_j) = a_{m_j}\left(X_{\left(\left[n - \frac{n}{m_j}\right],n\right)} - b_{m_j}\right) = a_{m_j}\left(\frac{\psi^{-1}(\tau_{r_j,n})}{a_n} + b_n - b_{m_j}\right),$$

or equivalently,

$$\tau_{r_j,n} = \psi\left(\frac{a_n}{a_{m_j}}W_n(\lambda_j) - a_n(b_n - b_{m_j})\right). \tag{37}$$

This implies that conditioning $\{S_n(\cdot,\lambda_j) : 1 \le j \le k\}$ on $\{W_n(\lambda_j) : 1 \le j \le k\}$ is equivalent to conditioning on $\{\tau_{r_j,n} : 1 \le j \le k\}$. According to the situation (i.e. (vM)(1),(2) or (3)) we fix $\sigma_j = \sigma_{\lambda_j}$ and $\mu_j = \mu_{\lambda_j}$ as given. Note again that,

$$S(x,\lambda_j) = N'(\psi(\sigma_j x - \mu_j)) \tag{38}$$

and

$$\tau_{r_j} = \psi(\sigma_j W(\lambda_j) + \mu_j). \tag{39}$$

From (36), (37), (38) and (39) it follows that to complete the argument we need only to check that the weak limit of the conditional distribution of $N_n(\cdot)$ given $W_n(\lambda_j) = u_j$ for $j = 1, \ldots, k$ is that of $N'(\cdot|\underline{u})$ or equivalently that the conditional distribution of $N_n(\cdot)$ given $\tau_{r_{j,n}} = \psi\left(\frac{a_n}{a_{m_j}} u_j - a_n(b_n - b_{m_j})\right)$ converges to that of $N'(\cdot|\underline{u})$.

We start with the latter. Given $\tau_{r_j} = \psi(\sigma_j u_j + \mu_j)$ for $1 \leq j \leq k$, the remaining $n-k$ jump points of $N'(x)$, $0 \leq x \leq \tau_{r_k}$ are distributed as the concatenation of the following k blocks of order statistics of independent samples: the first block correspond to the first $r_1 - 1$ jump points, which are distributed like the order statistics of a sample of size $r_1 - 1$ from the $U(0, \tau_{r_1})$ distribution; the second block correspond to the $r_2 - r_1 - 1$ jump points between τ_{r_1} and τ_{r_2} and are distributed like the order statistics of a sample of size $r_2 - r_1 - 1$ from $U(\tau_{r_1}, \tau_{r_2})$ and so on until the kth block which correspond to the $r_k - r_{k-1} - 1$ jump points between $\tau_{r_{k-1}}$ and τ_{r_k} and are distributed like the order statistics of a sample of size $r_k - r_{k-1} - 1$ from $U(\tau_{r_{k-1}}, \tau_{r_k})$. Finally, for $x > \tau_{r_k}$, $N'(x) - N'(x - \tau_{r_k})$ is again a standard Poisson process.

We now consider the conditional distribution of $N_n(\cdot)$ (or S_n) given $\tau_{r_j,n}$. From an obvious extension of Theorem 2.7 of David [1981], it follows that given order statistics, $Y_{i_1}, \ldots, Y_{i_k}$ of a sample $Y_1, \ldots, Y_n$ iid from a density $g \equiv G'$, the remaining $n-k$ order statistics are distributed like the $k+1$ blocks of order statistics of independent samples: $Y_1^{(1)}, \ldots, Y_{i_1-1}^{(1)}$ from density g_1; $Y_1^{(2)}, \ldots, Y_{i_2-i_1-1}^{(2)}$ from density g_2; and so on until $Y_1^{(k+1)}, \ldots, Y_{n-i_k}^{(k+1)}$ from density g_{k+1}

$$\begin{aligned}
g_1(y) &= g(y)1\left(y < Y_{(i_1)}\right)/G\left(Y_{(i_1)}\right); \\
g_j(y) &= g(y)1\left(Y_{(i_{j-1})} < y < Y_{(i_j)}\right)/\left(G\left(Y_{i_j}\right) - G\left(Y_{i_{j-1}}\right)\right), \\
&\qquad j = 2, \ldots, k; \\
g_{k+1}(y) &= g(y)1\left(y > Y_{(i_k)}\right)/\left(1 - G\left(Y_{(i_k)}\right)\right).
\end{aligned}$$

Now, for $1 \leq j \leq k$ let $Y_{in}^{(j)}$ be iid with conditional distribution functions (conditioned on $\tau_{r_j,n}$):

$$G_n^{(1)}(x) = \frac{\bar{F}\left(X_{(n-r_1+1,n)}\right) - \bar{F}\left(\frac{x}{a_n} + b_n\right)}{\bar{F}\left(X_{(n-r_1+1,n)}\right)}, \quad x > \psi^{-1}(\tau_{r_1,n}),$$

$$
\begin{aligned}
G_n^{(j)}(x) &= \frac{\bar{F}\left(X_{(n-r_{j-1}+1,n)}\right) - \bar{F}\left(\frac{x}{a_n}+b_n\right)}{\bar{F}\left(X_{(n-r_{j-1}+1,n)}\right) - \bar{F}\left(X_{(n-r_j+1,n)}\right)}, \\
&\qquad\qquad \psi^{-1}(\tau_{r_j,n}) < x < \psi^{-1}(\tau_{r_{j-1},n}),\ \ j = 2,\ldots,k, \\
G_n^{(k+1)}(x) &= \frac{F\left(\frac{x}{a_n}+b_n\right)}{F\left(X_{(n-r_k+1,n)}\right)}, \quad x < \psi^{-1}(\tau_{r_k,n}).
\end{aligned}
$$

From (33) it follows that,

$$
\begin{aligned}
G_n^{(1)}(x) &\sim 1 - \frac{\psi(x)}{\tau_{r_1,n}} \\
G_n^{(j)}(x) &\sim \frac{\tau_{r_{j-1},n} - \psi(x)}{\tau_{r_{j-1},n} - \tau_{r_j,n}} \qquad 1 \le j \le k \\
G_n^{(k+1)} &\sim F\left(\frac{x}{a_n}+b_n\right).
\end{aligned}
$$

Set,

$$
\begin{aligned}
S_n^{(1)}(x) &= \sum_{i=1}^{r_1} 1\left(Y_{in}^{(1)} > x\right), \\
S_n^{(j)}(x) &= \sum_{i=1}^{r_j - r_{j-1} - 1} 1\left(Y_{in}^{(j)} > x\right), \qquad 2 \le j \le k, \\
S_n^{(k+1)}(x) &= \sum_{i=1}^{n - r_k - 1} 1\left(Y_{in}^{(k+1)} > x\right).
\end{aligned}
$$

Combining the above remarks with this set-up, we see that given $\tau_{r_j,n}$ for $1 \le j \le k$, $S_n(x) \equiv S_n(x,1)$ is distributed as the concatenation of $k+1$ independent processes, $S_n^{(j)}(\cdot)$,

$$
S_n(x) = \begin{cases} S_n^{(1)}(x), & x > \psi^{-1}(\tau_{r_1,n}) \\ S_n^{(j)}(x), & \psi^{-1}(\tau_{r_j,n}) < x < \psi(\tau_{r_{j-1},n}), \quad 2 \le j \le k \\ S_n^{(k+1)}(x), & x < \psi^{-1}(\tau_{r_k,n}) \end{cases}.
$$

It follows that if $\tau_{r_j} = \psi(\sigma_j u_j - \mu_j)$ then given $\tau_{r_j,n} = \psi(u_j a_n/a_{m_j} - a_n(b_n - b_{m_j}))$ for $1 \le j \le k$, $G_{j,n}$ converges weakly to G_j, where

$$
G_j(x) = \frac{1 - \frac{\psi(x)}{\tau_{r_j}}}{1 - \frac{\psi(\tau_{r_j})}{\psi(\tau_{r_{j-1}})}}.
$$

Therefore, $N_n^{(j)}(x)$ for $0 < x < \psi(\sigma_k u_k + \mu_k)$ has the limiting distribution of $N'(\cdot, \underline{u})$, since $G_j\left(\psi^{-1}(\cdot)\right)$ is the Uniform distribution on $\tau_{r_{j-1}} < x < \tau_{r_j}$.

Finally, since

$$n\bar{F}\left(a_n\psi^{-1}(x) + b_n\right) \to \psi\psi^{-1}(x) = x,$$

we can show that

$$N_n^{(k+1)}(\cdot) - N_n(\tau_{r_k,n}) = \sum_{j=1}^{n} 1\left(X_j > \frac{\psi^{-1}(x)}{a_n} + b_n\right) - r_k, \quad x > \tau_{r_k,n},$$

converges weakly to a standard Poisson process, by simply checking finite dimensional joint distributions.

Part (b) of the agenda and conclusion (ii) of the lemma follows from the above and the usual Poisson convergence theorem.

□

(Part (ii) of the lemma is needed for Lemma 5).

Lemma 5 :

(i) Define

$$r \equiv r(\lambda) = \begin{cases} \left[\frac{1}{\lambda}\right] + 1, & \text{if } \frac{1}{\lambda} \text{ is not an integer} \\ \frac{1}{\lambda}, & \text{otherwise} \end{cases}.$$

then the marginal distribution of $\tilde{S}(x,\lambda) + r(\lambda)$ is Poisson with parameter $\sigma(\lambda)x$.

(ii) Suppose $U(\lambda,\cdot)$ is the limit law of $U_n(\lambda)$ and $U(\lambda,\cdot)$ has the distribution specified in Lemma 3, that is $U(\lambda,x)$ has the distribution of $\exp(-\lambda\tilde{S}(x,\lambda))$. Reparametrize $U(\lambda,\cdot)$ by r given above, say $U(\lambda,\cdot) = \tilde{U}(r,\cdot)$. Then, $r \to \tilde{U}(r,\cdot)$ is $1-1$, $r = 1, 2, \ldots$.

Proof . Note again that for fixed $x, \lambda = 1/m$, $\tilde{S}(x,\lambda)$ given $W(\lambda) = u$ is distributed as $N'(\psi(\sigma(\lambda)(x+u)+\mu(\lambda)))$ given $\tau_r = \psi(\sigma(\lambda)u+\mu(\lambda)$. But, $N'(\psi(\sigma(\lambda)(x+u)+\mu(\lambda))) - N'(\psi(\sigma(\lambda)u+\mu(\lambda)))$ is independent of $N'(\psi(\sigma(\lambda)u + \mu(\lambda))) = r$ given $\tau_r = \psi(\sigma(\lambda)u + \mu(\lambda))$ and has a Poisson $\psi(\sigma(\lambda)x)$ distribution and claim (i) follows.

Therefore,

$$\mathrm{E}\left(e^{-\lambda \tilde{S}(x,\lambda)}\right) = e^{-\lambda r} \exp\left(\psi(\sigma(\lambda)x) \cdot \left(e^{-\lambda} - 1\right)\right).$$

For (vM)(3), $\psi(u) = e^{-u}$ and $\sigma(\lambda) = 1$. Therefore, if $r \to \tilde{U}(r)$ is not 1-1 then the function $\lambda \to \exp\{-r(\lambda)\lambda\} \exp\{e^{-x}(e^{-\lambda} - 1)\}$ is not 1-1 as a map from $(0, 1)$ to functions of x. But this is evidently false. The argument for (vM)(1) and (vM)(2) is the same for the function $\lambda \to exp\{-\lambda r(\lambda)\} \exp\{\exp\{(x/\lambda^2)^\beta\}(\exp\{-\lambda\} - 1)\}$ for $\beta = \pm\gamma$.

Comment: This weakening of (A.5) suffices since $\hat{m}$ is obtained by a search over $m_j = [q^j n]$ where $\left[\frac{n}{m_j}\right]$ ranges over distinct integers. □

Peter J. Bickel
Department of Statistics,
University of California,
Berkeley, CA, 94720-3860
E-mail: bickel@stat.berkeley.edu

Anat Sakov
Department of Statistics and Operations Research,
School of Mathematical Sciences,
Tel Aviv University,
Ramat Aviv,
Tel Aviv 69978, ISRAEL
E-mail: sakov@post.tau.ac.il

Table 1: Coverage for b_n, known a_n

	Distribution	Exponential			Normal			Gumbel		
	n	500	1000	10,000	500	1000	10,000	500	1000	10,000
	Coverage	0.914	0.907	0.935	0.949	0.934	0.926	0.916	0.914	0.933
	Mean	16	15	34	23	27	64	15	15	27
$\hat{m}$	SD	11	9	25	15	16	45	9	10	21
	Median	12	14	24	16	24	57	12	11	24

	Distribution	Uniform		
	n	500	1000	10,000
	Coverage	0.93	0.926	0.945
	Mean	16	16	32
$\hat{m}$	SD	12	13	25
	Median	12	11	24

Table 2: Coverage for b_n, unknown a_n

	Distribution	Exponential			Normal			Gumbel		
	n	500	1000	10,000	500	1000	10,000	500	1000	10,000
	Coverage	0.84	0.85	0.88	0.82	0.86	0.88	0.8	0.84	0.88
	Mean	17	18	40	17	19	40	16	17	33
$\hat{m}$	SD	10	12	28	10	12	31	9	11	26
	Median	16	14	32	16	14	32	12	14	24
$\hat{\alpha}$	Mean	-0.02	-0.01	0	0.18	0.16	0.13	0.02	0.01	0
	SD	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

	Distribution	Uniform		
	n	500	1000	10,000
	Coverage	0.89	0.92	0.92
	Mean	17	19	39
$\hat{m}$	SD	12	12	28
	Median	16	18	32
$\hat{\alpha}$	Mean	0.94	0.94	0.98
	SD	0.1	0.1	0.1

Table 3: Three other cases

	Case	μ, known SD				μ, unknown SD			θ		
	Distribution	Exponential		Normal		Normal			Uniform		
	n	100	500	100	500	100	500	1000	100	500	1000
	Coverage	0.922	0.93	0.953	0.961	0.952	0.951	0.947	0.912	0.926	0.936
	Mean	54	195	37	132	58	188	315	10	14	15
$\hat{m}$	SD	29	156	29	149	27	154	298	6	8	8
	Median	57	159	24	67	57	119	178	8	12	11

Figure 1: Choice of m using the rule.