Biometrika (2010), **97**, 4, *pp*. 881–892
doi: 10.1093/biomet/asq045
Advance Access publication 4 September 2010

Bootstrap confidence intervals and hypothesis tests for extrema of parameters

BY PETER HALL AND HUGH MILLER

Department of Mathematics and Statistics, University of Melbourne, Melbourne, Victoria 3010, Australia

halpstat@ms.unimelb.edu.au h.miller@ms.unimelb.edu.au

SUMMARY

The bootstrap provides effective and accurate methodology for a wide variety of statistical problems which might not otherwise enjoy practicable solutions. However, there still exist important problems where standard bootstrap estimators are not consistent, and where alternative approaches, for example the m-out-of-n bootstrap and asymptotic methods, also face significant challenges. One of these is the problem of constructing confidence intervals or hypothesis tests for extrema of parameters, for example for the maximum of p parameters where each has to be estimated from data. In the present paper we suggest approaches to solving this problem. We use the bootstrap to construct an accurate estimator of the joint distribution of centred parameter estimators, and we base the procedure, either a confidence interval or a hypothesis test, on that distribution estimator. Our methodology is designed so that it errs on the side of conservatism, modulo the small inaccuracy of the bootstrap step.

Some key words: Bootstrap consistency; Conservative method; Coverage accuracy; Level accuracy; Maxima; Minima; Subsampling; Tie.

1. INTRODUCTION

Bootstrap methods can face serious difficulties when used to estimate the distributions of extrema of parameter estimators, for example of $\max(\hat\theta_1, \ldots, \hat\theta_p)$ where $\hat\theta_1, \ldots, \hat\theta_p$ are estimators of the respective values of parameters $\theta_1, \ldots, \theta_p$. The reason is that even the asymptotic distribution of $\max_j \hat\theta_j$ can be nonnormally distributed. A bootstrap meta-theorem argues that, in a range of settings, bootstrap methods give consistent results for estimating distributions of parameter estimators if and only if the limiting distribution is normal (see e.g. Mammen, 1992). Consequently, since the joint distribution of $\hat\theta_1 - \theta_1, \ldots, \hat\theta_p - \theta_p$ generally is asymptotically normal, that distribution can typically be estimated accurately even though the limiting distribution of $\max_{j \leqslant p} \hat\theta_j$ might not be estimable by any method, be it the bootstrap or another approach. This property underpins the methodology introduced in the present paper.

It might be thought that the m-out-of-n bootstrap (see e.g. Swanepoel, 1986; Hall, 1990; Bickel & Ren, 1996; Bickel et al., 1997; Politis et al., 1999) could overcome the problems faced by the standard bootstrap in estimating the distribution of quantities such as $\max_{j \leqslant p} \hat\theta_j$. However, that is not always the case. Andrews (2000) notes this difficulty, and it will be explored further in § 3. Even in problems where the m-out-of-n bootstrap enjoys attractive asymptotic properties, it can exhibit very poor finite-sample performance because the noise introduced through estimating the tuning parameter, m, in the m-out-of-n bootstrap can seriously degrade performance. As

a result, the m-out-of-n bootstrap can produce confidence intervals and hypothesis tests with serious anticonservative level inaccuracies.

Although these challenges stand in the way of accurate statistical methodology, the problem of making inference about extrema of parameters is important because it arises in a variety of contexts, in fields ranging from frontier analysis (see Berger & Humphrey, 1997; Kim et al., 2007) to methodology based on empirical eigenvalues (e.g. Ringrose & Benn, 1997; Schott, 2006). Contributors to the area include Beran (1982), Bretagnolle (1983), Beran & Srivastava (1985), Dümbgen (1993), Hall et al. (1993), Andrews (1999, 2000), Xie et al. (2009) and Hall & Miller (2009). Nevertheless, there still do not exist methods that overcome effectively the difficulties discussed above.

In the present paper we show that, in a variety of problems involving hypothesis tests and confidence intervals for extrema of general linear combinations of parameters, an indirect application of the bootstrap can overcome the difficulties. Our approach involves implicitly constructing a bootstrap confidence interval for the joint distribution of $\hat\theta_1 - \theta_1, \ldots, \hat\theta_p - \theta_p$, and using simple monotonicity arguments to construct tests or confidence intervals that are guaranteed to be conservative except for a small bootstrap error. We suggest using a double bootstrap approach to ensure good coverage accuracy. The conservatism of our methodology derives from the fact that our bootstrap methods, for both confidence intervals and hypothesis tests, are based on resampling from the least favourable null distribution, interpreted in a nonparametric context.

An interesting feature of the problems treated in this paper is that, in the cases that cause most difficulty, the distributions that we seek to approximate are asymmetric even in the asymptotic limit. In consequence, even two-sided confidence regions and two-sided hypothesis tests have level errors of order $n^{-1/2}$, not n^{-1}. Use of the double bootstrap is necessary to ensure that the bootstrap error is of order n^{-1} in two-sided, as well as one-sided, cases.

In today's computing environment, in cases where the $\hat\theta_j$s are relatively simple functions of the data, and when only a single sample is involved, it is often feasible to use triple bootstrap methods and thereby reduce error to order $n^{-3/2}$. However, it would be impractical to explore the effectiveness of the triple bootstrap in a simulation study, since this would increase computational labour several hundred fold. In principle, analytical rather than bootstrap methods, based on theoretical expressions for high-order terms in asymptotic formulae, might be used to effect further corrections. However, we feel that the complexity of the formulae involved, and the fact that they change from one problem to another, make this type of correction unattractive.

Although the intervals and tests tend towards conservatism, they are constructed so that, in difficult cases where there are ties for unknown parameter values, the procedures are less conservative, and in some cases asymptotically exact, modulo the bootstrap approximation. We explore the extent of this conservatism in our numerical work; it varies from being vanishingly small, when ties are present, to being more significant in other cases.

We suggest a general approach that enables us to construct conservative tests and confidence intervals for a very wide variety of statistics based on parameter extrema. The quantities that can be addressed using our methodology include confidence intervals for, and hypothesis tests about, functions of the parameters $\theta_1, \ldots, \theta_p$ such as the following:

$$\max_{1 \leqslant j \leqslant p} \theta_j, \quad \min_{1 \leqslant j \leqslant p} \theta_j, \quad \max_{j \in \mathcal{J}_1} \theta_j \pm \max_{j \in \mathcal{J}_2} \theta_j, \quad \min_{j \in \mathcal{J}_1} \theta_j \pm \max_{j \in \mathcal{J}_2} \theta_j,$$
$$\max\Big(\max_{1 \leqslant j \leqslant p} \theta_j, C \Big), \quad \max\Big(\min_{1 \leqslant j \leqslant p} \theta_j, C \Big), \tag{1}$$

where $\mathcal{J}_1$ and $\mathcal{J}_2$ are arbitrary subsets of $1, \ldots, p$, and C is any known constant. In practice $\mathcal{J}_1$ and $\mathcal{J}_2$ would usually be disjoint, but even this condition is not required for the general problem solved in § 2.

2. Methodology

2·1. *Problem setup*

First we give notation used throughout. Let $\mathcal{J}_1$ and $\mathcal{J}_2$ be nonempty subsets of $\{1, \ldots, p\}$, and, for $k = 1$ and 2, put $\max_{(k)} = \max_{j \in \mathcal{J}_k} \theta_j$ and $\hat{\max}_{(k)} = \max_{j \in \mathcal{J}_k} \hat{\theta}_j$. Define $\min_{(k)}$ and $\hat{\min}_{(k)}$ analogously. Our general treatment allows either extreme of each population to be considered, so write $\mathrm{m}_{(k)}$ to denote either $\min_{(k)}$ or $\max_{(k)}$, and define $\hat{\mathrm{m}}_{(k)}$ to equal $\hat{\max}_{(k)}$ if we took $\mathrm{m}_{(k)}$ to equal $\max_{(k)}$, and to equal $\hat{\min}_{(k)}$ otherwise. We wish to test H_0 against H_1, where

$$H_0 : \mathrm{m}_{(1)} \geqslant \mathrm{m}_{(2)}, \quad H_1 : \mathrm{m}_{(1)} < \mathrm{m}_{(2)}, \tag{2}$$

and again the test errs on the conservative side; or we wish to construct a confidence interval with the property:

$$\mathrm{pr}\{\mathrm{m}_{(2)} - \mathrm{m}_{(1)} \in [\hat{\mathrm{m}}_{(2)} - \hat{\mathrm{m}}_{(1)} - c_\alpha, \infty)\} \geqslant 1 - \alpha.$$

In practice, c_α will not be known and must be estimated. For this step we use the bootstrap to compute an empirical approximation, $\hat{c}_\alpha$ say, to c_α. The error committed here is quite low, particularly if the double bootstrap is employed, since our methodology ensures that c_α is defined in terms of the joint distribution of the centred variables $\hat{\theta}_j - \theta_j$; that distribution is relatively accessible.

Our discussion above, and our development of methodology below, focuses on one-sided conservative procedures. However, there is no difficulty extending our methodology to two-sided approaches in the usual way, by taking the intersection of two one-sided procedures.

2·2. *Obtaining conservative tests*

Recall the definition of $\mathrm{m}_{(k)}$ in § 2·1. Write $\mathrm{m}_{j \in \mathcal{J}_k} \theta_j$ for $\min_{j \in \mathcal{J}_k} \theta_k$ or $\max_{j \in \mathcal{J}_k} \theta_j$, according to whether $\mathrm{m}_{(k)}$ denotes the former or latter. Extend that notation to $\mathrm{m}_{j \in \mathcal{J}_k} x_j$ for a general quantity x_j, and observe the following:

$$\begin{aligned}
&\mathrm{pr}\big(\mathrm{m}_{(2)} - \mathrm{m}_{(1)} + c_\alpha \geqslant \hat{\mathrm{m}}_{(2)} - \hat{\mathrm{m}}_{(1)}\big) \\
&\quad = \mathrm{pr}\Big\{\mathrm{m}_{(2)} - \mathrm{m}_{(1)} + c_\alpha \geqslant \mathop{\mathrm{m}}_{j \in \mathcal{J}_2} (\hat{\theta}_j - \theta_j + \theta_j) - \mathop{\mathrm{m}}_{j \in \mathcal{J}_1} (\hat{\theta}_j - \theta_j + \theta_j)\Big\} \\
&\quad \geqslant \mathrm{pr}\Big[\mathrm{m}_{(2)} - \mathrm{m}_{(1)} + c_\alpha \geqslant \mathop{\mathrm{m}}_{j \in \mathcal{J}_2} \Big\{ \max_{k \in \mathcal{J}_2} (\hat{\theta}_k - \theta_k) + \theta_j \Big\} \\
&\qquad - \mathop{\mathrm{m}}_{j \in \mathcal{J}_1} \Big\{ \min_{k \in \mathcal{J}_1} (\hat{\theta}_k - \theta_k) + \theta_j \Big\}\Big] \\
&\quad = \mathrm{pr}\Big\{c_\alpha \geqslant \max_{k \in \mathcal{J}_2} (\hat{\theta}_k - \theta_k) - \min_{k \in \mathcal{J}_1} (\hat{\theta}_k - \theta_k)\Big\} = 1 - \alpha,
\end{aligned} \tag{3}$$

where the final identity holds provided that we define c_α by

$$\mathrm{pr}\Big\{ \min_{k \in \mathcal{J}_1} (\hat{\theta}_k - \theta_k) - \max_{k \in \mathcal{J}_2} (\hat{\theta}_k - \theta_k) \geqslant -c_\alpha \Big\} = 1 - \alpha.$$

Hence, if we take $\hat{c}_\alpha$ to be an empirical approximation to c_α, then, no matter what our choice of min and max in the definitions of $\mathrm{m}_{(1)}$ and $\mathrm{m}_{(2)}$, the confidence interval

$$[\hat{\mathrm{m}}_{(2)} - \hat{\mathrm{m}}_{(1)} - \hat{c}_\alpha, \infty)$$

covers $\mathrm{m}_{(2)} - \mathrm{m}_{(1)}$ with probability at least $1 - \alpha$, modulo any error in the approximation of c_α by $\hat{c}_\alpha$. Analogously, the hypothesis test that rejects $H_0 : \mathrm{m}_{(1)} \geqslant \mathrm{m}_{(2)}$, in favour of $H_1 : \mathrm{m}_{(1)} < \mathrm{m}_{(2)}$,

Table 1. *Possible hypothesis tests of extremes, along with corresponding confidence intervals and equations for obtaining* c_α

Case	H_0	H_1	Confidence Interval
1	$\min_{(1)} \geqslant C$	$\min_{(1)} < C$	$\min_{(1)} \in (-\infty, \hat{\min}_{(1)} + c_\alpha]$
2	$\min_{(1)} \leqslant C$	$\min_{(1)} > C$	$\min_{(1)} \in [\hat{\min}_{(1)} - c_\alpha, \infty)$
3	$\max_{(1)} \geqslant C$	$\max_{(1)} < C$	$\max_{(1)} \in (-\infty, \hat{\max}_{(1)} + c_\alpha]$
4	$\max_{(1)} \leqslant C$	$\max_{(1)} > C$	$\max_{(1)} \in [\hat{\max}_{(1)} - c_\alpha, \infty)$
5	$\min_{(1)} \geqslant \max_{(2)}$	$\min_{(1)} < \max_{(2)}$	$\max_{(2)} - \min_{(1)} \in [\hat{\max}_{(2)} - \hat{\min}_{(1)} - c_\alpha, \infty)$
6	$\min_{(1)} \leqslant \max_{(2)}$	$\min_{(1)} > \max_{(2)}$	$\min_{(1)} - \max_{(2)} \in [\hat{\min}_{(1)} - \hat{\max}_{(2)} - c_\alpha, \infty)$
7	$\max_{(1)} \geqslant \max_{(2)}$	$\max_{(1)} < \max_{(2)}$	$\max_{(2)} - \max_{(1)} \in [\hat{\max}_{(2)} - \hat{\min}_{(1)} - c_\alpha, \infty)$
8	$\min_{(1)} \geqslant \min_{(2)}$	$\min_{(1)} < \min_{(2)}$	$\min_{(2)} - \min_{(1)} \in [\hat{\max}_{(2)} - \hat{\min}_{(1)} - c_\alpha, \infty)$

Case	Equation for c_α
1	$\mathrm{pr}\{\min_{k\in\mathcal{J}_1}(\hat\theta_k - \theta_k) \geqslant -c_\alpha\} = 1-\alpha$
2	$\mathrm{pr}\{\max_{k\in\mathcal{J}_1}(\hat\theta_k - \theta_k) \leqslant c_\alpha\} = 1-\alpha$
3	$\mathrm{pr}\{\min_{k\in\mathcal{J}_1}(\hat\theta_k - \theta_k) \geqslant -c_\alpha\} = 1-\alpha$
4	$\mathrm{pr}\{\max_{k\in\mathcal{J}_1}(\hat\theta_k - \theta_k) \leqslant c_\alpha\} = 1-\alpha$
5	$\mathrm{pr}\{\max_{k\in\mathcal{J}_2}(\hat\theta_k - \theta_k) - \min_{k\in\mathcal{J}_1}(\hat\theta_k - \theta_k) \leqslant c_\alpha\} = 1-\alpha$
6	$\mathrm{pr}\{\max_{k\in\mathcal{J}_1}(\hat\theta_k - \theta_k) - \min_{k\in\mathcal{J}_2}(\hat\theta_k - \theta_k) \leqslant c_\alpha\} = 1-\alpha$
7	$\mathrm{pr}\{\max_{k\in\mathcal{J}_2}(\hat\theta_k - \theta_k) - \min_{k\in\mathcal{J}_1}(\hat\theta_k - \theta_k) \leqslant c_\alpha\} = 1-\alpha$
8	$\mathrm{pr}\{\max_{k\in\mathcal{J}_2}(\hat\theta_k - \theta_k) - \min_{k\in\mathcal{J}_1}(\hat\theta_k - \theta_k) \leqslant c_\alpha\} = 1-\alpha$

if $\hat{\mathrm{m}}_{(2)} - \hat{\mathrm{m}}_{(1)} - \hat{c}_\alpha > 0$, has level at most α, except for the error in the bootstrap approximation:

$$\begin{aligned}
&\mathrm{pr}\big(\hat{\mathrm{m}}_{(2)} - \hat{\mathrm{m}}_{(1)} - \hat{c}_\alpha > 0 \,\big|\, \mathrm{m}_{(2)} \leqslant \mathrm{m}_{(1)}\big) \\
&\quad \leqslant \mathrm{pr}\big(\hat{\mathrm{m}}_{(2)} - \hat{\mathrm{m}}_{(1)} - \hat{c}_\alpha > \mathrm{m}_{(2)} - \mathrm{m}_{(1)} \,\big|\, \mathrm{m}_{(2)} \leqslant \mathrm{m}_{(1)}\big) \\
&\quad = \mathrm{pr}\big\{\mathrm{m}_{(2)} - \mathrm{m}_{(1)} \notin \big[\hat{\mathrm{m}}_{(2)} - \hat{\mathrm{m}}_{(1)} - \hat{c}_\alpha, \infty\big) \,\big|\, \mathrm{m}_{(2)} \leqslant \mathrm{m}_{(1)}\big\} \\
&\quad \leqslant 1 - (1-\alpha) = \alpha, \qquad (4)
\end{aligned}$$

where the inequality in (3) follows from (4).

It is worth mentioning again that our focus on confidence intervals and hypothesis tests based on the difference $\mathrm{m}_{(2)} - \mathrm{m}_{(1)}$ does not restrict us to quantities such as $\max_{j\in\mathcal{J}_2}\theta_j - \min_{j\in\mathcal{J}_1}\theta_j$. By taking one or more of the components of θ to be null we can treat any of the quantities in (1), and others, in the way discussed above. To make this explicit, Table 1 shows a broad range of hypotheses that may be treated, along with the corresponding confidence interval form and the equation defining c_α. As noted before, conservative two-sided confidence intervals may be created by intersection. In the case of ties, and when $\mathrm{m}_{(1)}$ and $\mathrm{m}_{(2)}$ denote the minimum and maximum, respectively, the inequality in the formulae at (3) is in fact an equality; that is, the test is exact. This implies that cases 1, 4 and 5 in Table 1 are potentially exact. In other cases, where $\mathrm{m}_{(2)}$ and $\mathrm{m}_{(1)}$ are not max and min, respectively, the tied case is also conservative, and this conservatism will tend to grow with the number of populations under consideration.

3. APPROXIMATING DISTRIBUTIONS OF EXTREMA OF ESTIMATORS

3·1. *Models, and the challenges of distribution approximations*

For specificity in this section we treat the case where the quantity of interest is $\min_{j\in\mathcal{J}_1} - \max_{j\in\mathcal{J}_2}$, where $\mathcal{J}_1 = \{1,\ldots,r\}$, $\mathcal{J}_2 = \{r+1,\ldots,p\}$ and $r = 1,\ldots,p-1$. In regular problems the estimators $\hat\theta_j$ are root-n consistent and asymptotically normally distributed,

where n denotes sample size. Therefore it is reasonable to suppose that we can write

$$\hat{\theta}_j = \theta_j + n^{-1/2} N_j \qquad (j = 1, \ldots, p),$$

where $N_1, \ldots, N_p$ have a joint limiting normal distribution with zero mean. In this context, in § 3·2 we explore properties of

$$\hat{\omega} \equiv \min_{j \in \mathcal{J}_1} \hat{\theta}_j - \max_{j \in \mathcal{J}_2} \hat{\theta}_j = \min_{j \in \mathcal{J}_1} (\theta_j + n^{-1/2} N_j) - \max_{j \in \mathcal{J}_2} (\theta_j + n^{-1/2} N_j). \tag{5}$$

In the present paper the challenges of distribution approximation arise even under very generous assumptions, for example if $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)$ is the mean of a sample of size n from a normal $N(\theta, \Sigma)$ distribution where Σ is known. We shall show that, even in this case, difficulties with near ties can make it impossible to estimate consistently the asymptotic distribution of $\hat{\omega}$, at (5), no matter whether we use the bootstrap or any other method.

To give a simple example, assume that

$$\theta_2 = \theta_1 + n^{-1/2} \nu \text{ and } \theta_j = \theta_1 \text{ for } j = 3, \ldots, p, \text{ where } \nu \text{ is a fixed constant.} \tag{6}$$

Standard information-theoretic arguments show that ν is not identifiable, in the sense that it cannot be estimated consistently from data. Now, (5) and (6) imply that

$$n^{1/2} \hat{\omega} = \min(N_1, N_2 + \nu, N_3, \ldots, N_r) - \max(N_{r+1}, \ldots, N_p),$$

where, when $\hat{\theta}$ is the estimated mean of an n-sample from a normal $N(\theta, \Sigma)$ population, $(N_1, \ldots, N_p)$ is distributed as normal $N(0, \Sigma)$. The shape of the distribution of $n^{1/2} \hat{\omega}$, not just its location and scale, depend in detail on the nonestimable quantity ν.

The null hypothesis H_0, at (2), holds if $\nu \geqslant 0$, and the alternative obtains if $\nu < 0$. A conventional approach to constructing an asymptotically exact test of H_0 would be to estimate the null distribution of $\hat{\omega}$ for $\nu \geqslant 0$, and to reject H_0 if the value of $\hat{\omega}$ were less than an estimator of a lower critical point for the distribution of $\hat{\omega}$. However, since the distribution of $\hat{\omega}$ cannot be estimated consistently then this approach is not viable.

3·2. *Using the bootstrap to estimate the distribution of the centred version of* $\hat{\omega}$

Recall the definition of $\hat{\omega}$ at (5). In the case of the example suggested in § 3·1 the centred version of $\hat{\omega}$ with which we work is

$$\tilde{\omega} \equiv \min_{j \in \mathcal{J}_1} (\hat{\theta}_j - \theta_j) - \max_{j \in \mathcal{J}_2} (\hat{\theta}_j - \theta_j).$$

If $\hat{\theta}_j$ is constructed from a random sample vector $\mathcal{X} = \{X_1, \ldots, X_n\}$, and if $\mathcal{X}^* = \{X_1^*, \ldots, X_n^*\}$ denotes a resample drawn from $\mathcal{X}$ by sampling randomly, with replacement, then we write $\hat{\theta}_j^*$ for the version of $\hat{\theta}_j$ computed from $\mathcal{X}^*$ rather than $\mathcal{X}$. The bootstrap form of $\tilde{\omega}$ is

$$\tilde{\omega}^* = \min_{j \in \mathcal{J}_1} (\hat{\theta}_j^* - \hat{\theta}_j) - \max_{j \in \mathcal{J}_2} (\hat{\theta}_j^* - \hat{\theta}_j).$$

A percentile-bootstrap estimator of the distribution function F of $\tilde{\omega}$, defined by $F(x) = \text{pr}(\tilde{\omega} \leqslant x)$, is $\hat{F}(x) = \text{pr}(\tilde{\omega}^* \leqslant x \mid \mathcal{X})$.

The theoretical critical point c_α, corresponding to the confidence interval of case 5 in Table 1, is defined by $F(-c_\alpha) = \alpha$. We could define its estimator, $\hat{c}_\alpha$, simply as the single bootstrap solution, $c = \tilde{c}_\alpha$ say, of $\hat{F}(-c) = \alpha$. However, a greater degree of accuracy is obtained by using the double bootstrap. The aim is to find a $\beta \in [0, 1]$ such that $\hat{F}(-\tilde{c}_\beta)$ targets α more accurately. Given a resample $\mathcal{X}^*$, we draw, with replacement, a re-resample denoted by $\mathcal{X}^{**} = \{X_1^{**}, \ldots, X_n^{**}\}$, from which we compute the corresponding version of $\tilde{\omega}$, denoted by $\tilde{\omega}^{**}$. Put

$\hat{F}^*(x) = \mathrm{pr}(\widetilde{\omega}^{**} \leqslant x \mid \mathcal{X}^*)$, and let $c = \tilde{c}^*_\alpha$ be the solution of $\hat{F}^*(-c) = \alpha$. Let $\beta = \hat{\beta}(\alpha)$ be the solution of $\mathrm{pr}(\hat{\omega}^* \leqslant -\tilde{c}^*_\beta \mid \mathcal{X}) = \alpha$. In this notation the double-bootstrap estimator of c_α is $\hat{c}_\alpha = \tilde{c}_{\hat{\beta}(\alpha)}$. Of course, in practice the probabilities $\mathrm{pr}(\widetilde{\omega}^{**} \leqslant x \mid \mathcal{X}^*)$ and $\mathrm{pr}(\hat{\omega}^* \leqslant -\tilde{c}^*_\beta \mid \mathcal{X})$ usually cannot be computed exactly and are instead calculated by simulating many versions of $\mathcal{X}^*$ and $\mathcal{X}^{**}$. Thus to summarize, c is found by first determining a β such that the βth percentile of $\widetilde{\omega}^{**}$ is greater than the corresponding $\widetilde{\omega}^*$ with probability α, with this probability estimated by averaging over all single bootstrap resamples, and then adopting the βth percentile of $\widetilde{\omega}^*$, to arrive at the refined critical point.

3·3. *Accuracy of the bootstrap*

In the Appendix we show that the single-bootstrap critical point $\tilde{c}_\alpha$ satisfies

$$\tilde{c}_\alpha = c_\alpha + O_p(n^{-1}), \tag{7}$$

as $n \to \infty$. The fact that the error here is n^{-1}, rather than $n^{-1/2}$, reflects only the fact that we have not normalized when defining c_α. Indeed, in asymptotic terms,

$$n^{1/2}\, c_\alpha = u(\alpha) + O\bigl(n^{-1/2}\bigr), \tag{8}$$

where $u(\alpha)$ is defined in terms of a p-variate normal distribution; see (A7).

The $O_p(n^{-1})$ error in (7) is comprised primarily of errors incurred in estimating the variance matrix, and in fact (7) can be written in more detail as

$$\tilde{c}_\alpha = c_\alpha + n^{-1/2}\, (\hat{\Sigma} - \Sigma)^{\mathrm{T}}\, w + O_p\bigl(n^{-3/2}\bigr), \tag{9}$$

where $\hat{\Sigma}$ denotes the bootstrap estimator of the variance matrix of the p-vector $n^{1/2}\,(\hat{\theta} - \theta)$, and is approximated implicitly in the process of calculating $\tilde{c}_\alpha$; and w is a fixed vector of length equal to the number of components of Σ. Result (9) is derived in the Appendix. There we also outline a proof that the coverage error of a confidence interval, or level error of a hypothesis test, if we use the single-bootstrap critical point $\tilde{c}_\alpha$, rather than its double-bootstrap counterpart, is of size $n^{-1/2}$, not n^{-1}. We show too that this level of accuracy prevails in both one- and two-sided procedures; the parity properties that are familiar in more conventional cases do not, in this context, lead to a reduction to $O(n^{-1})$ accuracy in two-sided problems.

4. NUMERICAL PROPERTIES

4·1. *University rankings*

The relative ranking of universities continues to attract interest both within academia and in the broader community. The methodology suggested in this paper allows comparisons to be made between groups of universities. For example, suppose we wanted to explore whether Switzerland or the Netherlands had, in some sense, the strongest university for science. One statistic of interest could be the average number of articles published in the journals *Science* or *Nature* per year. This is one of the pieces of information used in the popular Shanghai Jiao Tong University rankings; see www.arwu.org. Figure 1(a) shows the distribution of the number of papers for two of the leading universities in Switzerland and the top four universities in the Netherlands for the 12 years from 1997 to 2008. Another leading Swiss university, ETH Zurich, has not been considered here due to its nonstationarity over this time period; for instance, it had five papers in both 2003 and 2004, while it produced 20 and 37 in 2007 and 2008 respectively. Aside from this omission, a university was included in the comparison if it had the most number of papers published for that country in an individual year.

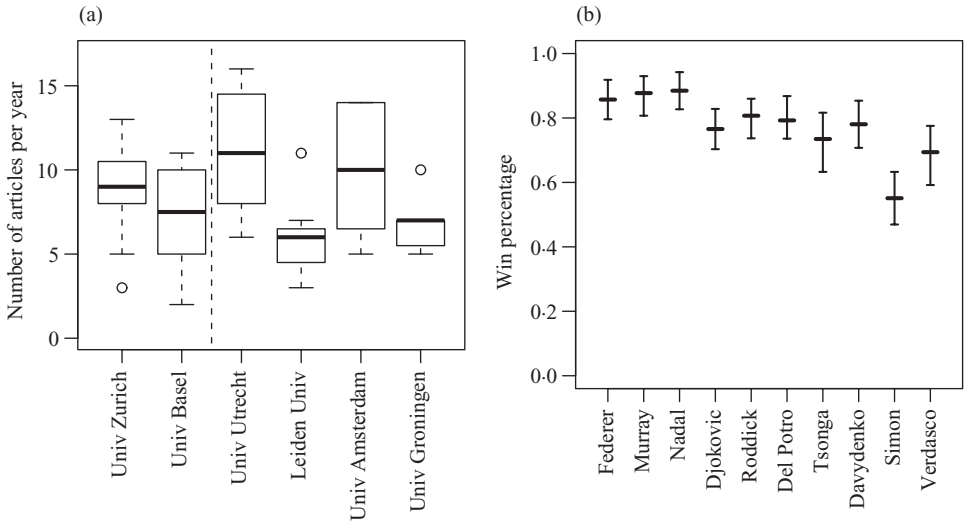

Fig. 1. Comparison of universities and of male tennis players. (a) Contains boxplots of the number of articles published per year in *Science* or *Nature* for Swiss and Dutch institutions. (b) Shows the winning percentages for the world top ten male tennis players.

Our statistic $\hat{\theta}_j$ is the mean number of papers published over the 12 years. Utrecht University in the Netherlands has the highest overall mean, so we use the test described in case 7 of Table 1 with Swiss universities as the first population and Dutch universities comprising the second. The double bootstrap with $B = 999$ resamples in each layer resulted in a p-value of 0·14, suggesting that there is moderate evidence for the Netherlands having the stronger university in the particular sense defined above.

The most important observation is that this significance test is more appropriate than pairwise comparisons between universities, since it recognizes that we are not entirely sure which institution is the best performing for each country. For instance, if we calculated a p-value in a similar fashion but only compared the University of Zurich and Utrecht University, having the observed maximum mean for Switzerland and The Netherlands, respectively, we would obtain a value of 0·054. The significance level is misleadingly high because we have ignored the role of other universities. While our test does not guarantee maximum power, it does give a better indication of certainty.

We make a few further comments about our results. First, the comparison statistic is obviously somewhat simplified. We pay no attention to whether or not the author from a given institution is listed first or otherwise, any relationships between articles and any changes over the 12-year periods. Many of these issues are common to similar analyses. Secondly, this particular example is interesting in our context because of uncertainty regarding the strongest university from each country. If a similar study was to be completed comparing the USA to Japan, say, a pairwise test without conservatism would be appropriate, since each country has a clear leader, Harvard University and the University of Tokyo, respectively, for mean number of papers published.

4·2. *Tennis player performance*

Figure 1(b) shows the winning proportion of the top 10 ranked male tennis players; that is, the number of matches won divided by the number of matches played. In the figure the players are ordered according to their official ranking, current at 20 August 2009. The winning proportion was calculated based on matches of the Association of Tennis Professionals in the 2009 calendar

Table 2. *Estimated p-values for the hypothesis that Simon's winning rate is as good as the minimum of the top t players, excluding himself*

t	1	2	3	4	5	6	7	8	9	10
p-value	0·000	0·001	0·001	0·019	0·024	0·027	0·059	0·082	N/A	0·214

year up to the same date. Eighty percent confidence intervals for these percentages are included in the figure as well. The most notable feature is that Simon, ranked 9, has a much lower winning percentage than the other players. Using the test of case 6 from Table 1, we can find p-values under the null hypothesis that Simon's performance is at least as good as that of the worst of the other top t players. Table 2 shows the p-values, estimated by means of the double bootstrap with $B = 999$ resamples in each layer, for various choices of t. The case $t = 9$ is irrelevant since Simon himself occupies that ranking. The results suggest there is weak evidence, $p = 0{\cdot}21$, that Simon is below everyone else in the top 10, but increasingly strong evidence for smaller values of t. For instance, the corresponding p-value comparing Simon to the most inferior of the top six players is 0·027. The bootstrap resampling step was conducted independently, and so ignored any dependence introduced by players having matches against each other.

In this example it is important to test the multiple hypothesis, since it is not at all clear who has the lowest winning percentage after Simon. Apart from the top three players, the remaining six have heavily overlapping confidence intervals.

4·3. *Wisconsin breast cancer*

The final real data example is included to demonstrate the utility of this approach in cases where conservatism is not an issue, and to introduce other possible statistics, besides the mean, which may be of interest. The Wisconsin dataset was first introduced by Wolberg & Mangasarian (1990) and is downloadable from the UCI machine learning database, http://archive.ics.uci.edu/ml/index.html. It is comprised of 699 observations, each with nine variables addressing tumour characteristics, along with an assignment of malignancy. While the main emphasis is usually the prediction of malignancy, an important part of any model is ensuring that relationships among predictors are understood. Therefore we focus here on determining which two of the nine predictor variables have the highest pairwise Pearson correlation. The correlation statistic is asymptotically normal, suggesting that our bootstrap methodology is appropriate. Variables two and three have the highest pairwise correlation by a good margin: 0·91 compared to the next highest, 0·76. We test the hypothesis that the correlation between variables two and three is larger than for any of the other 35 pairs, using the test from case 7 in Table 1. The resulting p-value is less than 0·001. Thus we can apply the method easily to problems that might otherwise be difficult to treat formally.

4·4. *Simulation of conservatism*

The following example shows the increasing conservatism as the differences among the true means θ_j diverge. Suppose we have $p = 10$ populations and the parameter of interest is the mean, which takes values $\{t, 2t, \ldots, pt\}$ in the populations, for some scalar t. We are interested in constructing an upper confidence interval for the maximum of these, as in case 4 in Table 1. We assume the underlying observations are standard normal and each population has n observations from which to estimate the mean. In this example we can find c_α analytically, allowing us to better focus on the degree of conservatism, rather than on estimation error. We know that the $(1-\alpha)$th quantile, d_α say, of the distribution of the maximum of p standard normal random variables is given by $F_p\{\Phi(d_\alpha)\} = 1-\alpha$, where $F_p(x) = x^p$ for $0 \leqslant x \leqslant 1$ and Φ is the standard

Table 3. *Simulated coverage percentages exploring conservatism in* § 4·4

	t					
n	0	0·2	0·4	0·6	0·8	1
10	90·1 (0·2)	94·6 (0·2)	96·5 (0·1)	97·7 (0·1)	98·1 (0·1)	98·3 (0·1)
20	89·9 (0·2)	95·8 (0·1)	97·6 (0·1)	98·1 (0·1)	98·5 (0·1)	98·5 (0·1)
50	90·2 (0·2)	96·8 (0·1)	98·3 (0·1)	98·6 (0·1)	98·7 (0·1)	98·9 (0·1)
100	90·1 (0·2)	97·5 (0·1)	98·5 (0·1)	98·6 (0·1)	98·8 (0·1)	99·0 (0·1)

Table 4. *Simulated coverage percentages for the example in* § 4·4 *with additional initial hypothesis test*

	t					
n	0	0·2	0·4	0·6	0·8	1
10	89·7 (0·2)	94·3 (0·2)	96·1 (0·1)	97·3 (0·1)	97·4 (0·1)	97·4 (0·1)
20	89·6 (0·2)	95·5 (0·1)	97·2 (0·1)	97·4 (0·1)	97·4 (0·1)	97·1 (0·1)
50	89·8 (0·2)	96·4 (0·1)	97·4 (0·1)	97·3 (0·1)	97·1 (0·1)	96·8 (0·1)
100	89·7 (0·2)	97·1 (0·1)	97·4 (0·1)	96·8 (0·1)	96·6 (0·1)	96·1 (0·1)

normal cumulative distribution function. Thus, for each simulation we may set $c_\alpha = n^{-1/2} d_\alpha$. Table 3 shows the estimated coverage probabilities over 20 000 simulations for the $1 - \alpha = 90\%$ confidence interval for various choices of t and n. Standard errors are shown in brackets.

Increasing conservatism is evident as we move across the table from the left, where there is no conservatism, to the right. Another evident feature is that if anything, conservatism increases with the number of observations. This is perhaps not unsurprising, since by treating all the observed $\hat{\theta}_j$s as equal in the hypothesis test we lose any benefit associated with increased sample size.

The issue of conservatism being independent of sample size can be addressed by adding a preceding step to our analysis. For a given simulation and $k < p$, we can perform a test of the hypothesis that the maximum of the k populations with smallest means is below the maximum of the other $p - k$, as in the seventh line of Table 1. If we find the maximum k such that we reject the null at some suitably low level, say $\alpha = 0{\cdot}02$, then we can construct a confidence interval for the overall maximum using only the remaining $p - k$ populations. The results of this approach are presented in Table 4.

The table shows that for a fractional loss of conservatism when all observations are in fact tied, in the first column of the table, we have significantly reduced the conservatism in situations where the $\hat{\theta}_j$ are well separated. In fact, as either t or n grows, the coverage again approaches the target value 0·9.

4·5. *Illustration of the accuracy of the double bootstrap*

We give two illustrative examples comparing the coverage accuracy of the double bootstrap to that for the single bootstrap, where the αth percentile of the bootstrapped means is used, and the normal approximation, where the estimated mean is assumed to follow its asymptotic t-distribution. In the first the means are sampled from the exponential distribution with mean 1 and we test the coverage of one-sided 80% confidence intervals. The second has means sampled from a Pareto distribution with the mean equal to 2, a scale parameter of 1 and shape parameter of 2, tested at 90% confidence. In each case $p = 10$ and we tested a range of n. We used $B = 599$ resamples for each bootstrap layer, and averaged over 2000 simulations. Results are presented in Table 5. In the exponential case, the double bootstrap enjoys good coverage accuracy for all n, with results lying within natural variation levels around 0·80. The single bootstrap underestimates the interval width, although gives reasonable results for $n \geqslant 20$. The normal approximation overestimates the interval width, and this effect persists for n of moderate size. In the Pareto case all approaches

Table 5. *Simulated coverage percentages. Targeted coverage was* 80% *and* 90% *for the exponential and Pareto distributions, respectively*

	Exponential distribution			Pareto distribution		
	Single	Normal	Double	Single	Normal	Double
n	Bootstrap	Approximation	Bootstrap	Bootstrap	Approximation	Bootstrap
10	75·0 (1·0)	85·8 (0·8)	79·7 (0·9)	95·5 (0·5)	98·1 (0·3)	91·3 (0·6)
15	78·0 (0·9)	84·0 (0·8)	80·7 (0·9)	96·0 (0·4)	96·7 (0·4)	91·6 (0·6)
20	79·2 (0·9)	82·1 (0·9)	80·4 (0·9)	96·4 (0·4)	96·9 (0·4)	90·9 (0·6)
25	80·5 (0·9)	82·2 (0·9)	81·2 (0·9)	96·8 (0·4)	96·7 (0·4)	90·9 (0·6)

overestimate the confidence interval width, with the double bootstrap clearly preferred at all n tested.

The double bootstrap is usually computationally manageable; in the university and tennis examples the computation time was 38·3 and 14·5 seconds, respectively, on a standard desktop computer with an Intel Core 2 Duo processor. Thus, if the dataset is sufficiently large and well behaved then the user may find the accuracy of the single bootstrap or normal approximation sufficient, but otherwise the double bootstrap is recommended over competing approaches.

ACKNOWLEDGEMENT

The research of both authors was partly funded by the Australian Research Council.

APPENDIX

Technical details for § 3

Under the smooth-function model (Bhattacharya & Ghosh, 1978) an estimator, or a vector of estimators such as $\hat\theta = (\hat\theta_1, \ldots, \hat\theta_p)$, computed from a sample of size n, is represented as a smooth function of a mean of n independent random vectors all distributed as V, say. If the distribution of V has sufficiently many finite moments and satisfies Cramér's condition, i.e. $\limsup_{\|t\|\to\infty} |E\{\exp(it^{\mathrm{T}}V)\}| < 1$, for which it is sufficient that the distribution of V be nonsingular; if the function has sufficiently many derivatives; and if the limiting variance-covariance matrix of $\hat\theta$, Σ say, is nonsingular; then, for $r \geqslant 1$,

$$\mathrm{pr}\{n^{1/2}(\hat\theta_j - \theta_j) \leqslant z_j,\ 1 \leqslant j \leqslant p\} = \Phi_\Sigma(z) + \sum_{k=1}^{r} n^{-k/2} P_k(z)\,\phi_\Sigma(z) + O\bigl(n^{-(r+1)/2}\bigr), \qquad \text{(A1)}$$

uniformly in $z = (z_1, \ldots, z_p)$, where ϕ_Σ and Φ_Σ are, respectively, the density and distribution functions of the normal $N(0, \Sigma)$ distribution, $P_1, \ldots, P_k$ are polynomials, not depending on n, with coefficients depending on derivatives of the smooth functions in the smooth-function model, evaluated at moments of V. The number of moments required of V increases with r. See Bhattacharya & Ranga Rao (1976) and Bhattacharya & Ghosh (1978). Analogously, (A1) has an empirical version

$$\begin{aligned}\mathrm{pr}\{n^{1/2}(\hat\theta_j^* - \hat\theta_j) \leqslant z_j \text{ for } 1 \leqslant j \leqslant p \mid \mathcal{X}\} \\ = \Phi_{\hat\Sigma}(z) + \sum_{k=1}^{r} n^{-k/2} \hat P_k(z)\,\phi_{\hat\Sigma}(z) + O_p\bigl(n^{-(r+1)/2}\bigr), \qquad \text{(A2)}\end{aligned}$$

where $\mathcal{X}$ denotes the dataset from which $\hat\theta_1, \ldots, \hat\theta_p$ were computed, $\hat\Sigma$ is the bootstrap estimator of Σ calculated from $\mathcal{X}$, and $\hat P_k$ is the version of P_k when moments of V, appearing in P_k, are replaced by their empirical counterparts. See, for example, chapter 3 of Hall (1992).

Next we discuss the true critical point c_α and the single bootstrap to estimator, $\tilde c_\alpha$, defined by

$$\mathrm{pr}\Bigl\{\max_{j\in\mathcal{J}_2}(\hat\theta_j - \theta_j) - \min_{j\in\mathcal{J}_1}(\hat\theta_j - \theta_j) \leqslant c_\alpha\Bigr\} = 1 - \alpha \qquad \text{(A3)}$$

and

$$\mathrm{pr}\Big\{\max_{j\in\mathcal{J}_2}(\hat\theta_j^*-\hat\theta_j)-\min_{j\in\mathcal{J}_1}(\hat\theta_j^*-\hat\theta_j)\leqslant\tilde c_\alpha\;\Big|\;\mathcal{X}\Big\}=1-\alpha, \tag{A4}$$

respectively. Put $d_\alpha=n^{1/2}\,c_\alpha$, $\tilde d_\alpha=n^{1/2}\,\tilde c_\alpha$, $Z_j=n^{1/2}\,(\hat\theta_j-\theta_j)$ and $Z_j^*=n^{1/2}\,(\hat\theta_j^*-\hat\theta_j)$. In this notation, (A1)–(A4) imply that

$$\begin{aligned}1-\alpha&=\mathrm{pr}\Big\{\max_{j\in\mathcal{J}_2}Z_j-\min_{j\in\mathcal{J}_1}Z_j\leqslant d_\alpha\Big\}\\&=\int_{\mathcal{A}(d_\alpha)}\Big[\phi_\Sigma(z)+n^{-1/2}\,\frac{d}{dz}\,\{P_1(z)\,\phi_\Sigma(z)\}\Big]\,dz+O(n^{-1}),\end{aligned} \tag{A5}$$

$$\begin{aligned}1-\alpha&=\mathrm{pr}\Big\{\max_{j\in\mathcal{J}_2}Z_j^*-\min_{j\in\mathcal{J}_1}Z_j^*\leqslant\tilde d_\alpha\;\Big|\;\mathcal{X}\Big\}\\&=\int_{\mathcal{A}(\tilde d_\alpha)}\Big[\phi_{\hat\Sigma}(z)+n^{-1/2}\,\frac{d}{dz}\,\big\{\hat P_1(z)\,\phi_{\hat\Sigma}(z)\big\}\Big]\,dz+O_p(n^{-1}),\end{aligned} \tag{A6}$$

where $\mathcal{A}(d)$ is the set of $z\in\mathbb{R}^p$ such that $z_{j_2}-z_{j_1}\leqslant d$ for all $j_1\in\mathcal{J}_1$ and all $j_2\in\mathcal{J}_2$, and we have taken $r=1$ in (A1) and (A2). In this notation $u(\alpha)$, at (8), is the solution of the equation

$$1-\alpha=\int_{\mathcal{A}\{u(\alpha)\}}\phi_\Sigma(z)\,dz. \tag{A7}$$

Define $e_\alpha=e_\alpha(\Sigma)$ to be the solution of the equation $\int_{\mathcal{A}(e_\alpha)}\phi_\Sigma(z)\,dz=1-\alpha$. If Σ_1 is a general nonsingular covariance matrix then

$$|e_\alpha(\Sigma_1)-e_\alpha(\Sigma)-v(\Sigma_1-\Sigma)^{\mathrm{T}}\,\dot e_\alpha(\Sigma)|\leqslant C_1\,\|\Sigma_1-\Sigma\|^2, \tag{A8}$$

whenever $\|\Sigma_1-\Sigma\|\leqslant C_2$, where $v(M)$ is the vector of length $\frac{1}{2}\,p\,(p+1)$ defined as a concatination of the distinct components of a general symmetric matrix M, $\dot e_\alpha(\Sigma)$ is the vector of derivatives of $e_\alpha(\Sigma)$ with respect to the components of $v(\Sigma)$, $\|\cdot\|$ denotes any given matrix norm and C_1 and C_2 are positive constants depending only on Σ. In view of (A5), $d_\alpha=e_{1-\beta}(\Sigma)+O(n^{-1})$, where

$$\beta=1-\alpha-n^{-1/2}\int_{\mathcal{A}(d_\alpha)}\frac{d}{dz}\,\{P_1(z)\,\phi_\Sigma(z)\}\,dz.$$

Since $(d/dz)\,\{\hat P_1(z)-P_1(z)\}\,\phi_\Sigma(z)=O_p(n^{-1/2})$, uniformly in z, and $\hat\Sigma=\Sigma+O_p(n^{-1/2})$, then (A6) yields $\tilde d_\alpha=e_{1-\beta}(\hat\Sigma)+O_p(n^{-1})$. Hence, by (A8) and the fact that $\dot e_{1-\beta}(\hat\Sigma)=\dot e_{1-\beta}(\Sigma)+O(n^{-1/2})$, we have $\tilde d_\alpha=d_\alpha+v(\hat\Sigma-\Sigma)^{\mathrm{T}}\,\dot e_{1-\beta}(\Sigma)+O_p(n^{-1})$. This then gives rise to (9), since

$$\tilde c_\alpha=n^{-1/2}\,c_\alpha+n^{-1/2}\,v(\hat\Sigma-\Sigma)^{\mathrm{T}}\,\dot e_{1-\beta}(\Sigma)+O_p\big(n^{-3/2}\big). \tag{A9}$$

The coverage error in a confidence interval, or level error in a hypothesis test, that we incur when using the single-bootstrap approximation, $\tilde c_\alpha$, to c_α is the value we obtain when substituting $\tilde c_\alpha$ for c_α in (A4) and subtracting $1-\alpha$. That is, the error equals

$$\mathrm{pr}\Big\{\max_{j\in\mathcal{J}_2}Z_j-\min_{j\in\mathcal{J}_1}Z_j\leqslant n^{1/2}\,\tilde c_\alpha\Big\}-(1-\alpha).$$

Substituting for $\tilde c_\alpha$ using (A9), or equivalently (9), we deduce that

$$\mathrm{pr}\Big\{\max_{j\in\mathcal{J}_2}Z_j-\min_{j\in\mathcal{J}_1}Z_j\leqslant d_\alpha+v(\hat\Sigma-\Sigma)^{\mathrm{T}}\,\dot e_{1-\beta}(\Sigma)+O_p(n^{-1})\Big\}-(1-\alpha). \tag{A10}$$

Unsurprisingly, standard arguments (see e.g. Ch. 5 of Hall, 1992) show that the $O_p(n^{-1})$ inside the probability in (A10), if dropped, produces a remainder term of order n^{-1}, and likewise that the term in $v(\hat\Sigma-\Sigma)^{\mathrm{T}}\,\dot e_{1-\beta}(\Sigma)$, being exactly of size $n^{-1/2}$, if ignored gives a remainder of exact size $n^{-1/2}$. Therefore the coverage error or level error of the single-bootstrap procedure is genuinely of size $n^{-1/2}$.

Analogously to (A9), the critical point $\tilde{c}^*_\alpha$, introduced in § 3·3, is

$$\tilde{c}^*_\alpha = n^{-1/2}\,\tilde{c}_\alpha + n^{-1/2}\,v(\hat{\Sigma}^* - \hat{\Sigma})^{\mathrm{T}}\,\dot{e}_\alpha(\Sigma) + O_p\bigl(n^{-3/2}\bigr).$$

Here we have used the fact that, owing to the smoothness of $\dot{e}_\alpha(\Sigma)$ as a function of Σ, $\dot{e}_\alpha(\hat{\Sigma}) = \dot{e}_\alpha(\Sigma) + O_p(n^{-1/2})$. The double bootstrap correctly captures the main cause of error, arising from the difference $\hat{\Sigma} - \Sigma$ in (A9), in the single-bootstrap approximation. That is, the double bootstrap correctly captures the first-order terms that describe departures of size $n^{-1/2}$ from the limiting distribution. A rigorous proof of this result follows using standard arguments given in Ch. 5 of Hall (1992).

REFERENCES

ANDREWS, D. W. K. (1999). Estimation when a parameter is on a boundary. *Econometrica* **67**, 1341–83.

ANDREWS, D. W. K. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica* **68**, 399–405.

BERAN, R. (1982). Estimated sampling distributions: the bootstrap and competitors. *Ann. Statist.* **10**, 212–25.

BERAN, R. & SRIVASTAVA, M. S. (1985). Bootstrap tests and confidence regions for functions of a covariance matrix. *Ann. Statist.* **13**, 95–115.

BERGER, A. & HUMPHREY, D. (1997). Efficiency of financial institutions: international survey and directions for future research. *Eur. J. Operational Res.* **98**, 175–212.

BHATTACHARYA, R. N. & GHOSH, J. K. (1978). On the validity of the formal Edgeworth expansion. *Ann. Statist.* **6**, 434–51.

BHATTACHARYA, R. N. & RANGA RAO, R. (1976). *Normal approximation and asymptotic expansions*. Wiley Series in Probability and Mathematical Statistics. New York, London, Sydney: John Wiley & Sons.

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

BICKEL, P. J. & REN, J.-J. (1996). The m out of n bootstrap and goodness of fit tests with double censored data. In *Robust statistics, data analysis, and computer intensive methods (Schloss Thurnau, 1994)*, pp. 35–47, vol. 109 of *Lecture Notes in Statist.* New York: Springer.

BRETAGNOLLE, J. (1983). Lois limites du bootstrap de certaines fonctionnelles. *Ann. Inst. H. Poincaré Sect. B (N.S.)* **19**, 281–96.

DÜMBGEN, L. (1993). On nondifferentiable functions and the bootstrap. *Probab. Theory Rel. Fields* **95**, 125–40.

HALL, P. (1990). Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *J. Mult. Anal.* **32**, 177–203.

HALL, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer Series in Statistics. New York: Springer-Verlag.

HALL, P., HÄRDLE, W. & SIMAR, L. (1993). On the inconsistency of bootstrap distribution estimators. *Comp. Statist. Data Anal.* **16**, 11–8.

HALL, P. & MILLER, H. (2009). Using the bootstrap to quantify the authority of an empirical ranking. *Ann. Statist.* **37**, 3929–59.

KIM, M., KIM, Y. & SCHMIDT, P. (2007). On the accuracy of bootstrap confidence intervals for efficiency levels in stochastic frontier models with panel data. *J. Productivity Anal.* **28**, 165–81.

MAMMEN, E. (1992). *When does bootstrap work? Asymptotic results and simulations*, vol. 77 of *Statistics and Computing*. New York: Springer-Verlag.

POLITIS, D. N., ROMANO, J. P. & WOLF, M. (1999). *Subsampling*. Springer Series in Statistics. New York: Springer-Verlag.

RINGROSE, T. & BENN, D. (1997). Confidence regions for fabric shape diagrams. *J. Structural Geol.* **19**, 1527–36.

SCHOTT, J. (2006). A high-dimensional test for the equality of the smallest eigenvalues of a covariance matrix. *J. Mult. Anal.* **97**, 827–43.

SWANEPOEL, J. (1986). A note on proving that the(modified) bootstrap works. *Commun. Statist. A* **15**, 3193–203.

WOLBERG, W. & MANGASARIAN, O. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proc. Nat. Acad. Sci. USA* **87**, 9193–96.

XIE, M., SINGH, K. & ZHANG, C. (2009). Confidence intervals for population ranks in the presence of ties and near ties. *J. Am. Statist. Assoc.* **104**, 775–88.

[*Received November* 2009. *Revised April* 2010]