A note on methods of restoring consistency to the bootstrap

By RICHARD SAMWORTH

Statistical Laboratory, Centre for Mathematical Sciences, Wilberforce Road, Cambridge, CB3 0WB, UK.

r.j.samworth@statslab.cam.ac.uk

Summary

We consider the property of consistency and its relevance for determining the performance of the bootstrap. We analyse various parametric bootstrap approximations to the distributions of the Hodges and Stein estimators, whose behaviour is typical of that of superefficient estimators employed in wavelet regression, kernel density estimation and nonparametric curve fitting. Our results reveal not only some of the difficulties in selecting good modifications to the intuitive bootstrap, but also that inconsistent bootstrap approximations may perform better than consistent versions, even in large samples.

Some key words: Bootstrap inconsistency; Hodges estimator; m out of n bootstrap; Stein estimator; Superefficiency.

1. Introduction

Consistency is seen as the sine qua non for the bootstrap. Much recent research has focused on known cases of bootstrap inconsistency and on methods which restore consistency. As with all asymptotic statistical properties, however, it is important to assess its relevance to the finite samples which face the practitioner.

In this article we argue that the pointwise asymptotics of consistency can mask the finite-sample behaviour, and that inconsistent bootstrap estimators may in fact perform better than their consistent counterparts. We illustrate this point with reference to the parametric bootstrap and the Hodges and Stein estimators. In both

cases, we find that the intuitive, inconsistent, parametric bootstrap outperforms the consistent methods we consider, namely the m out of n bootstrap and a refined choice of parameter estimate.

2. The Hodges estimator

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables, each distributed according to $N(\theta, 1)$, and let $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$. The Hodges estimator is defined by

$$T_{n,H} = \begin{cases} b\bar{X}_n & \text{if } |\bar{X}_n| \leq n^{-1/4} \\ \bar{X}_n & \text{otherwise,} \end{cases}$$

where $b \in (0, 1)$. We wish to estimate the distribution $H_n(\theta)$ of $n^{1/2}(T_{n,H} - \theta)$, and consider the bootstrap approximation $H_n(\bar{X}_n)$. The risk of the Hodges estimator, given by $E\{n(T_{n,H} - \theta)^2\}$, converges to b^2 when $\theta = 0$ and to 1 otherwise (Lehmann 1998, p.442). Thus the sequence $(T_{n,H})$ is asymptotically superefficient at $\theta = 0$, and Beran (1997) shows that the intuitive n out of n bootstrap above is inconsistent at points of asymptotic superefficiency.

The Hodges estimator is studied in this form for its simplicity and tractability. Similar superefficient truncation estimators have been studied in wavelet regression, where estimates of wavelet coefficients are discarded if smaller in modulus than some threshold value. Further details can be found in, for example, an unpublished technical report by A. J. Canty, A. C. Davison, D. V. Hinkley and V. Ventura.

Denoting by $H_n(x, \theta)$ the distribution function corresponding to $H_n(\theta)$, we find that

$$H_n(x, \theta) = \begin{cases} \Phi(x) & \text{if } x < -n^{1/4} - n^{1/2}\theta \\ \Phi(-n^{1/4} - n^{1/2}\theta) & \text{if } -n^{1/4} - n^{1/2}\theta \leq x < -bn^{1/4} - n^{1/2}\theta \\ \Phi[b^{-1}\{x + (1-b)\theta n^{1/2}\}] & \text{if } -bn^{1/4} - n^{1/2}\theta \leq x < bn^{1/4} - n^{1/2}\theta \\ \Phi(n^{1/4} - n^{1/2}\theta) & \text{if } bn^{1/4} - n^{1/2}\theta \leq x < n^{1/4} - n^{1/2}\theta \\ \Phi(x) & \text{if } x \geq n^{1/4} - n^{1/2}\theta, \end{cases} \tag{1}$$

where $\Phi(\cdot)$ denotes the distribution function of a standard normal random variable. Thus $H_n(\theta)$ converges weakly for all $\theta \in \mathbb{R}$, with limiting distribution $H(\theta)$ which is $N(0, 1)$ when $\theta \neq 0$ and $N(0, b^2)$ when $\theta = 0$. The bootstrap distribution also

converges weakly, in probability, to $N(0,1)$ when $\theta \neq 0$. However, when $\theta = 0$, Beran (1982) shows that $H_n(\bar{X}_n)$ converges, as a random element of the space of probability measures on the real line metrised by weak convergence, to the random probability measure $N\{(b-1)Z, b^2\}$, where $Z \sim N(0,1)$. Hence the standard n out of n bootstrap is consistent if and only if $\theta \neq 0$.

One procedure which restores consistency in this context involves reducing the bootstrap resample size, an idea which dates back to Bretagnolle (1983). To be specific, it follows from Corollary 2.1(b) of Beran (1997) that, if $m = m_n$ is chosen so that $m \to \infty$ but $m = o(n)$, then $H_m(\bar{X}_n)$ is consistent for all $\theta \in \mathbb{R}$.

[Figure 1 about here.]

In Fig. 1, we compare the errors in the bootstrap approximations $H_m(\bar{X}_n)$ as estimators of $H_n(\theta)$ for $m = n^{1/2}$, $m = n^{3/4}$ and $m = n$. These values of m are understood to be rounded to the nearest integer. Comparisons are made using the Kolmogorov distance, that is the supremum metric on the corresponding distribution functions. This distance is averaged over 1000 realisations of $\bar{X}_n$.

It is particularly interesting to note that, although smaller choices of m do improve the bootstrap performance in a very small neighbourhood of $\theta = 0$, the improvements come at the expense of considerably worse performance outside this neighbourhood. Treated as a problem in decision theory, the minimax rule appears to be to choose $m = n$, and this would agree with the Bayes rule unless most of the mass of the prior were concentrated in a very small neighbourhood of $\theta = 0$.

We give here a heuristic explanation for the results observed. Write $m^{1/2}\bar{X}_n = m^{1/2}\theta + m^{1/2}n^{-1/2}Z$, where $Z \sim N(0,1)$. From (1), we see that the magnitude of the error in the bootstrap approximation depends on the absolute value of the difference between $n^{1/2}\theta$ and $m^{1/2}\theta + m^{1/2}n^{-1/2}Z$. If $|\theta| \ll n^{-1/2}$, then the random term in the error, $m^{1/2}n^{-1/2}Z$, dominates. The variance of this term increases as m increases relative to n, although it always has zero expectation. However, for larger values of $|\theta|$, the fixed error, $\theta(m^{1/2} - n^{1/2})$ is crucial. This is large in absolute value for small m relative to n, and decreases to zero as m increases to n.

We now investigate whether or not it is possible to retain the desirable characteristics of both methods by means of an empirical, data-driven choice of m. That is, if we let $m = f_n(|\bar{X}_n|)$, where $f_n : [0, \infty) \to \{1, \ldots, n\}$ is some suitably chosen non-decreasing function, can we achieve improved performance in a neighbourhood of $\theta = 0$ without loss elsewhere in the parameter space?

The resulting bootstrap approximation will be consistent if $f_n(x_n) \to \infty$ whenever $x_n = \theta + O(n^{-1/2})$ for some $\theta \neq 0$, and $f_n(x_n) = o(n)$ whenever $x_n = O(n^{-1/2})$. A simple class of possible choices of m is given by

$$m = \begin{cases} An^{\alpha} & \text{if } |\bar{X}_n| \leq Bn^{-\beta} \\ n & \text{if } |\bar{X}_n| > Bn^{-\beta}, \end{cases} \tag{2}$$

where $A, B > 0$, $\alpha \in (0, 1)$ and $\beta \in (0, 1/2)$. The fact that the bootstrap estimators $H_m(\bar{X}_n)$ in this class are consistent follows from Corollary 1.1 of Putter & van Zwet (1996) and the fact that $\mathrm{pr}\{m = o(n)\} \to 1$ as $n \to \infty$ when $\theta = 0$.

Numerical studies suggest that improved performance in a small neighbourhood of $\theta = 0$ can be achieved, but that once again this comes at the expense of worse performance outside this neighbourhood. Although the 'bad' neighbourhoods vanish in the limit as n tends to infinity, which ensures consistency, they remain a problem in finite samples. The problem occurs in the region, in this case where $|\theta| \approx Bn^{-\beta}$, in which the event $\{|\bar{X}_n| \leq Bn^{-\beta}\}$ has moderate probability. Considered as an attempt to estimate the optimal value $m_{\mathrm{opt}} = m_{\mathrm{opt}}(\theta)$, the rule in (2) is analogous to using the Hodges estimator as an estimator of θ, and suffers the same drawbacks. Of course, other more complicated empirical choices of m are possible, but the scope for improvement over the naive n out of n bootstrap appears minimal.

A further suggestion for restoring consistency, proposed by Putter & van Zwet (1996), involves a refined choice of parameter estimate in the bootstrap approximation: we replace $H_n(\bar{X}_n)$ by $H_n(\hat{\theta}_n)$ where $\hat{\theta}_n$ is chosen so that

(i) $\mathrm{pr}(\hat{\theta}_n = 0) \to 1$ as $n \to \infty$ when $\theta = 0$

(ii) $\mathrm{pr}(\hat{\theta}_n \neq 0) \to 1$ as $n \to \infty$ when $\theta \neq 0$.

The consistency of $H_n(\hat{\theta}_n)$ then follows again from Corollary 1.1 of Putter & van Zwet (1996). The authors themselves suggest an estimator from the following class:

$$\hat{\theta}_n = \begin{cases} 0 & \text{if } |\bar{X}_n| \leq Bn^{-\beta} \\ \bar{X}_n & \text{if } |\bar{X}_n| > Bn^{-\beta}, \end{cases}$$

where $B > 0$ and $\beta \in (0, 1/2)$. Note that, when $B = 1$ and $\beta = 1/4$, $\hat{\theta}_n$ is the Hodges estimator with $b = 0$. Once again, however, the improvements in the immediate vicinity of $\theta = 0$ are offset by severe losses elsewhere in the parameter space. Comparing the expression for $H_n(x, \theta)$ in (1) with the corresponding $H_n(x, \hat{\theta}_n)$, we see that, when $\theta \in (n^{-1/2}, Bn^{-\beta})$, it is likely that $n^{1/2}\hat{\theta}_n$ will be zero, whereas $n^{1/2}\theta$ is large. Thus $H_n(x, \hat{\theta}_n)$ will be a poor estimator of $H_n(x, \theta)$ in this region of the parameter space.

3. The Stein estimator

Now suppose that $X_1, \ldots, X_n$ are independent and identically distributed random vectors in $\mathbb{R}^k$, where $k \geq 4$. Each X_i has a k-variate normal distribution $N_k(\theta, I)$, with mean vector $\theta \in \mathbb{R}^k$ and identity covariance matrix. Write $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$, define $\mu : \mathbb{R}^k \to \mathbb{R}$ by $\mu(x) = k^{-1} \sum_{i=1}^k x_i$, and let e denote a k-vector of ones. The Stein estimator $T_{n,S}$ is defined by

$$T_{n,S} = \mu(\bar{X}_n)e + \left(1 - \frac{k-3}{n\|\bar{X}_n - \mu(\bar{X}_n)e\|^2}\right)\{\bar{X}_n - \mu(\bar{X}_n)e\}.$$

Thus the Stein estimator 'shrinks' each component of $\bar{X}_n$ towards the mean of the nk observations. It has found many practical uses arising from the well-known fact that, when θ is estimated with respect to quadratic loss, $T_{n,S}$ strictly dominates $\bar{X}_n$. More generally, the behaviour of the Stein estimator in this regular parametric setting is symptomatic of that of superefficient shrinkage estimators employed in more general problems such as kernel density estimation and nonparametric regression. There, the complexity of the parameter space allows far more severe forms of superefficiency (Brown et al., 1997).

One notable difference between the Hodges and Stein estimators is that the Stein estimator improves on $\bar{X}_n$ as an estimator of θ, in terms of mean squared error,

for every n, not just asymptotically. Le Cam (1953) showed that one-dimensional asymptotically superefficient estimators necessarily perform poorly in a neighbourhood of a point of asymptotic superefficiency.

We are interested in the distribution $H_n(\theta)$ of $n^{1/2}(T_{n,S} - \theta)$, and consider the bootstrap estimator $H_n(\bar{X}_n)$. When the components of θ are not all equal, $H_n(\theta)$ converges weakly to $N_k(0, I)$, and $H_n(\bar{X}_n)$ converges weakly in probability to the same limit. However, when the components of θ are all equal, $H_n(\theta)$ converges to the the probability measure $\pi(0)$, where for any $h \in \mathbb{R}^k$, we define $\pi(h)$ to be the distribution of

$$Z - \frac{(k-3)\{Z - \mu(Z)e + h - \mu(h)e\}}{\|Z - \mu(Z)e + h - \mu(h)e\|^2},$$

where $Z \sim N_k(0, I)$. On the other hand, $H_n(\bar{X}_n)$ converges, as a random element of the space of probability distributions on $\mathbb{R}^k$ metrised by weak convergence, to the random probability measure $\pi(Z')$, where $Z' \sim N_k(0, I)$ and is independent of Z (Beran, 1997). Thus the standard n out of n bootstrap is inconsistent when the components of θ are all equal.

The m out of n bootstrap again restores consistency throughout the parameter space provided that $m = m_n$ satisfies $m \to \infty$, and $m = o(n)$. To facilitate an empirical comparison of different choices of m, we use a stochastic approximation to the supremum metric on the space of distribution functions on $\mathbb{R}^k$, an idea first suggested by Beran & Millar (1986). The algorithm consists of the following steps.

Step 1. Generate independent $\bar{X}_{n,1}, \ldots, \bar{X}_{n,R} \sim N_k(\theta, n^{-1}I)$ and compute $\hat{H}_{n,R}(\theta)$, the empirical distribution of $n^{1/2}\{T_{n,S}(\bar{X}_{n,1}) - \theta\}, \ldots, n^{1/2}\{T_{n,S}(\bar{X}_{n,R}) - \theta\}$.

Step 2. For $i = 1, \ldots, B$, repeat Steps 3 and 4.

Step 3. Generate $\bar{Y}_{n,i} \sim N_k(\theta, n^{-1}I)$ and conditionally independent $\bar{X}^*_{m,1}, \ldots, \bar{X}^*_{m,R} \sim N_k(\bar{Y}_{n,i}, m^{-1}I)$ in order to compute $\hat{H}_{m,R}(\bar{Y}_{n,i})$, the empirical distribution of $m^{1/2}\{T_{m,S}(\bar{X}^*_{m,1}) - \bar{Y}_{n,i}\}, \ldots, m^{1/2}\{T_{m,S}(\bar{X}^*_{m,R}) - \bar{Y}_{n,i}\}$.

Step 4. Generate independent $y_1, \ldots, y_{q_R} \sim N_k(0, I)$, and compute

$$d_i = \max_{1 \le q \le q_R} |\hat{H}_{m,R}(y_q, \bar{Y}_{n,i}) - \hat{H}_{n,R}(y_q, \theta)|,$$

where $\hat{H}_{n,R}(x, \theta)$ and $\hat{H}_{m,R}(x, \bar{Y}_{n,i})$ are the distribution functions corresponding to $\hat{H}_{n,R}(\theta)$ and $\hat{H}_{m,R}(\bar{Y}_{n,i})$ respectively.

Step 5. Compute $\bar{d} = B^{-1} \sum_{i=1}^{B} d_i$.

[Figure 2 about here.]

In Fig. 2, we plot $\bar{d}$ as a function of $\lambda = n\|\theta - \mu(\theta)e\|^2$. Note that $\lambda = 0$ corresponds to all the components of θ being equal. Numerical studies show no qualitative change for different θ-directions. We find that improvements at $\lambda = 0$ are possible, but there is still a price to be paid in terms of poor performance for larger values of λ.

To explain these observations, let Z, Z' denote independent standard k-variate normal random variables, and let $T^*_{m,S} = T_{m,S}(\bar{X}^*_n)$. We can write

$$\begin{aligned} n^{1/2}(T_{n,S} - \theta) &= n^{1/2}(\bar{X}_n - \theta) - \frac{(k-3)[n^{1/2}\{\bar{X}_n - \mu(\bar{X}_n)e\}]}{\|n^{1/2}\{\bar{X}_n - \mu(\bar{X}_n)e\}\|^2} \\ &\sim Z - \frac{(k-3)[Z - \mu(Z)e + n^{1/2}\{\theta - \mu(\theta)e\}]}{\|Z - \mu(Z)e + n^{1/2}\{\theta - \mu(\theta)e\}\|^2} \end{aligned}$$

and

$$\begin{aligned} m^{1/2}(T^*_{m,S} - \bar{X}_n) &= m^{1/2}(\bar{X}^*_m - \bar{X}_n) - \frac{(k-3)[m^{1/2}\{\bar{X}^*_m - \mu(\bar{X}^*_m)e\}]}{\|m^{1/2}\{\bar{X}^*_m - \mu(\bar{X}^*_m)e\}\|^2} \\ &\sim Z' - \frac{(k-3)[Z' - \mu(Z')e + m^{1/2}\{Z - \mu(Z)e\}/n^{1/2} + m^{1/2}\{\theta - \mu(\theta)e\}]}{\left\|Z' - \mu(Z')e + m^{1/2}\{Z - \mu(Z)e\}/n^{1/2} + m^{1/2}\{\theta - \mu(\theta)e\}\right\|^2}. \end{aligned}$$

Comparing these expressions, we see that $H_m(\bar{X}_n)$ will be a good approximation to $H_n(\theta)$ when $m^{1/2}\{Z - \mu(Z)e\}/n^{1/2} + m^{1/2}\{\theta - \mu(\theta)e\}$ is close to $n^{1/2}\{\theta - \mu(\theta)e\}$. When λ is small, and in particular when $\lambda = 0$, the main error is likely to come from the random term, $m^{1/2}\{Z - \mu(Z)e\}/n^{1/2}$. Since we can choose m to tend to infinity as slowly as we like, we can make this error as small as we like, in probability.

However, when λ is large, the dominant part is the difference between the two non-random terms, $m^{1/2}\{\theta - \mu(\theta)e\}$ and $n^{1/2}\{\theta - \mu(\theta)e\}$. This decreases to zero as m increases towards n, and explains the poor performance of the reduced resample size methods for larger values of λ.

Analogues of the empirical rules for choosing m and the Putter & van Zwet method of restoring consistency also exist for this problem. For instance, the latter may be implemented with

$$\hat{\theta}_n = \begin{cases} \mu(\bar{X}_n)e & \text{if } \|\bar{X}_n - \mu(\bar{X}_n)e\| \leq Bn^{-\beta} \\ \bar{X}_n & \text{otherwise,} \end{cases}$$

where $B > 0$ and $\beta \in (0, 1/2)$, in which case the resulting bootstrap approximation $H_n(\hat{\theta}_n)$ is consistent for all $\theta \in \mathbb{R}^k$. Although numerical studies suggest it is possible to achieve minor improvements for a fixed n with a suitable choice of B, any choice of B will eventually be poor for sufficiently large n, because $n\|\bar{X}_n - \mu(\bar{X}_n)e\|^2$ has a noncentral chi-squared distribution with $(k-1)$ degrees of freedom and noncentrality parameter λ, so the event $\{\|\bar{X}_n - \mu(\bar{X}_n)e\| \leq Bn^{-\beta}\}$ has moderate probability when $\lambda \approx B^2 n^{1-2\beta}$. Thus the event $\{\hat{\theta}_n = \mu(\bar{X}_n)e\}$ is eventually probable, even for large λ, and $H_n(\hat{\theta}_n)$ will then perform poorly. Similar remarks apply to empirical choices of m in the m out of n bootstrap.

As in § 2, we see that the standard parametric bootstrap performs better than expected relative to its competitors, and the fixed-parameter asymptotics of consistency tell only part of the story.

Acknowledgement

I am extremely grateful to my Ph.D. supervisor, Alastair Young, for suggesting this investigation and for many subsequent helpful conversations. I am also grateful for the comments of an anonymous referee on a previous draft. This research is supported by a studentship, from the UK Engineering and Physical Sciences Research Council.

References

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

Beran, R. J. (1997). Diagnosing bootstrap success. *Ann. Inst. Statist. Math.* **49**, 1–24.

Beran, R. J. & Millar, P. W. (1986). Confidence sets for a multivariate distribution. *Ann. Statist.* **14**, 431–43.

Bretagnolle, J. (1983). Lois limites du bootstrap de certaines fonctionnelles. *Ann. Inst. Henri Poincaré* **19**, 281–96.

Brown, L. D., Low, M. G. & Zhao, L. H. (1997). Superefficiency in nonparametric function estimation. *Ann. Statist.* **25**, 2607–25.

LeCam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. *Univ. Calif. Pub. Statist.* **1**, 277–330.

Lehmann, E. L. (1998). *Theory of Point Estimation*, 2nd ed. New York: Springer-Verlag.

Putter, H. & van Zwet, W. R. (1996). Resampling: consistency of substitution estimators. *Ann. Statist.* **24**, 2297–318.

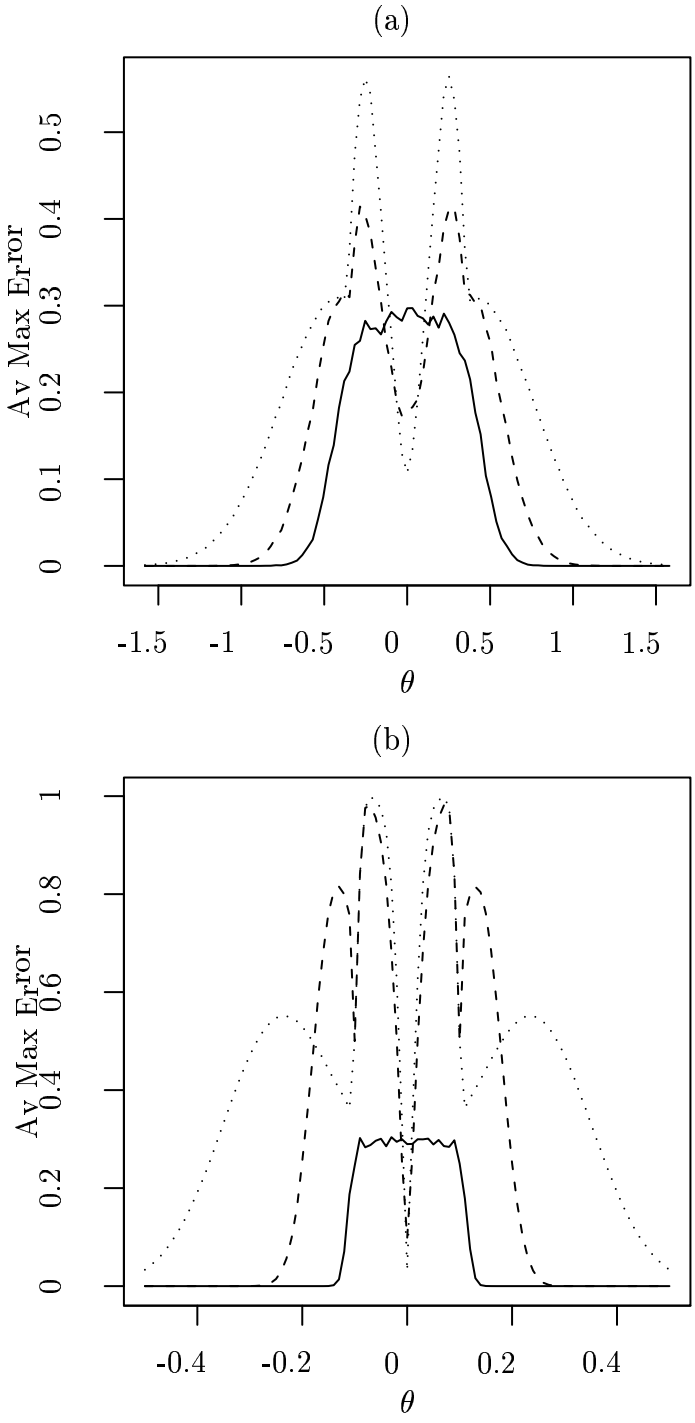

Figure 1: The average maximum error over 1000 realisations of the bootstrap approximations with $m_n = n^{1/2}$ (dotted), $m_n = n^{3/4}$ (dashed), $m_n = n$ (solid). Parameter values: $b = 0.5$; (a) $n = 100$; (b) $n = 10,000$.

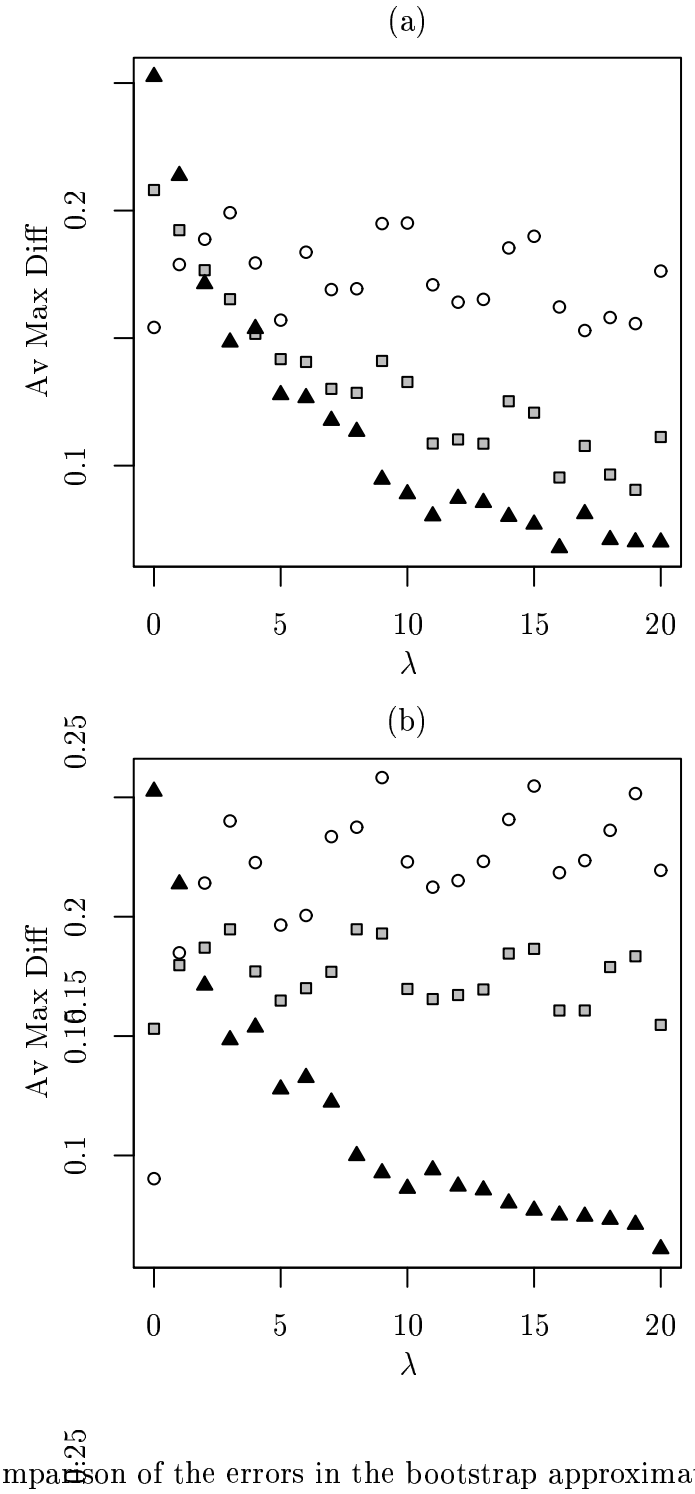

Figure 2: A comparison of the errors in the bootstrap approximations $\hat{H}_{m,R}(x, \bar{X}_n)$ for $m = n^{1/2}$ (circles), $m = n^{3/4}$ (grey squares) and $m = n$ (black triangles). Parameter values: $R = 500$; $q_R = 1,000$; $B = 200$; $k = 5$; $n^{1/2}\theta = (\lambda/2)^{1/2}(-1, 1, 0, 0, 0)$; (a) $n = 100$; (b) $n = 10,000$.