A resampling technique for massive data in settings of bootstrap inconsistency

Tim White

Submitted under the supervision of Dr. Snigdhansu Chatterjee to the University Honors Program at the University of Minnesota - Twin Cities in partial fulfillment of the requirements for the degree of Bachelor of Science, *summa cum laude* in Statistical Science.

May 10, 2022

Abstract

As massive data sets become more and more prevalent across the sciences, there is a growing need for accurate and computationally efficient methods of estimator quality assessment that can be applied under a variety of data-generating conditions. The nonparametric bootstrap is a straightforward and accurate method for approximating the sampling distribution of an estimator, but it becomes computationally unwieldy for large samples. Kleiner et al.'s bag of little bootstraps (BLB) provides a computationally efficient alternative to the bootstrap, but it is not expected to perform well in settings where the bootstrap is inconsistent. In this paper, we introduce the bag of little m out of n bootstraps (BLmnB), a modification of the BLB that aims to extend the method's applicability to cases of bootstrap inconsistency. We formalize the BLmnB algorithm and compare its performance against that of the bootstrap and the BLB in two well-documented settings of bootstrap failure. Our results indicate that while the BLmnB is capable of outperforming the other two methods in one of these settings, it performs no better than the BLB in the other. In both settings, we find that the approximation accuracy of the BLmnB is sensitive to the choice of the resample size m. While these findings suggest that the BLmnB is a promising alternative to the BLB in at least some data-generating scenarios, further investigation is necessary in order to develop the underlying theory of the method and study its accuracy and runtime in other settings of bootstrap inconsistency.

Contents

1 Introduction

1.1 Background and motivation

Estimator quality assessment is a fundamental component of statistical inference. Consider a scenario in which we observe data drawn independently and identically from some unknown population, and suppose that we are interested in some unknown parameter of this population. Suppose further that we have identified a statistic — i.e., some function of the data — that can be used as an estimator for our parameter of interest. This estimator is a random variable, and it has a distribution of possible values called the sampling distribution. It is common to use some summary of the sampling distribution — e.g., a confidence interval — as a measure of the uncertainty surrounding the estimator. However, the sampling distribution is unknown in practice, so it must be approximated.

The ideal frequentist approach to obtaining an assessment of estimator quality in this scenario would involve drawing many independent samples from the population, computing a realization of the estimator on each sample, forming the sampling distribution of the estimator from these many realizations, and directly computing the desired summary of the sampling distribution. Unfortunately, this approach is infeasible in practice since we generally do not have the time or resources to collect many independent and identically distributed samples. Often, we have access to just one sample.

In settings where the parameter of interest is the population mean, the Central Limit Theorem (CLT) provides a theoretical basis for approximating the sampling distribution from a single sample — it tells us that the sampling distribution of the sample mean is asymptotically normal even if the underlying population distribution is not normal (Hogg et al., 2019). However, there exists no analogous theoretical justification for approximating the sampling distribution from a single sample if our parameter of interest is something other than the population mean. The nonparametric bootstrap emerged in response to this lack of a widely applicable CLT equivalent, ushering in an era of computation-based resampling for the purpose of statistical inference.

1.2 A brief history of resampling

The nonparametric bootstrap (henceforth referred to as "the bootstrap") was proposed by Efron (1979) as a computationally intensive extension of an earlier resampling technique called the jackknife (Tukey, 1958). Efron's bootstrap accurately approximates the sampling distribution of nearly any estimator of interest by repeatedly sampling with replacement from a single observed sample, computing a realization of the estimator on each resample, and assembling the results into an empirical sampling distribution. We present this procedure in greater detail in Section 2.2.

The bootstrap has many favorable theoretical and practical qualities. It is generally consistent (Putter & van Zwet, 1996) and higher order correct (Hall, 1992). It is also a fairly automatic procedure, as one needs to select only one tuning parameter — the number of Monte Carlo iterations — to implement it. However, as noted by Kleiner et al. (2014), the bootstrap becomes computationally inefficient for large samples. For an observed sample of size n, each bootstrap resample of size n contains approximately $0.632n$ distinct data points from the original sample (Efron & Tibshirani, 1993). Thus, the repeated computation of estimates on a large number of bootstrap resamples requires considerable time and computing power in the massive data setting.

As the size of data sets has increased across the sciences and the computational constraints of the bootstrap have become more apparent, new computation-based inferential methods have emerged that achieve greater efficiency by operating on subsets of data that are smaller than the original sample. Subsampling (Politis et al., 1999) proposes generating many subsamples of size $b << n$ from a given sample of size n, computing a realization of the estimator of interest on each subsample, and assembling the results into an empirical sampling distribution. The m out of n bootstrap (Bickel et al., 1997) suggests modifying Efron's bootstrap algorithm by generating resamples of size $m < n$ instead of n. While these procedures require the repeated computation of estimates on only around $0.632b$ or $0.632m$ data points instead of $0.632n$, they are far less automatic than the bootstrap (Kleiner et al., 2014). For one thing, they are sensitive to the selection of additional tuning parameters — b for subsampling and m for the m out of n bootstrap. In addition, since both procedures rely on the computation of estimates on subsets containing fewer than n observations, they require knowledge of the convergence rate of the estimator of interest in order to recover the bootstrap's favorable theoretical properties.

A more recent innovation in resampling is the bag of little bootstraps (BLB), a method proposed by Kleiner et al. (2014) that combines the automatic nature and favorable theoretical properties of the bootstrap with the computational efficiency of subsampling and the m out of n bootstrap. We formalize the BLB algorithm in Section 2.3, but a general overview is as follows: (1) Form subsamples of size $b << n$ from a given sample of size n, (2) apply a multinomial vector of counts to each subsample — effectively sampling with replacement — to generate resamples of size n, (3) compute a realization of the estimator of interest for each resample, (4) assemble these estimates into an empirical sampling distribution and compute the desired estimator quality assessment for each subsample, and (5) aggregate these assessments across the subsamples. The advantages offered by this procedure stem from the second step — the BLB does not require knowledge of the estimator's convergence rate since the resamples of size n are on the scale of the original sample, and the computation of estimates on these resamples is relatively efficient since each resample contains at most b distinct data points. Hence, the BLB is generally a suitable computational method for conducting inference in the massive data setting.

While the BLB is applicable in all cases in which the bootstrap succeeds, Kleiner et al. (2014) note that it is not expected to perform well when the bootstrap fails. However, the authors suggest that a modified procedure — a so-called bag of little m out of n bootstraps (BLmnB), in which resamples of size $m < n$ are generated from each subsample in the BLB algorithm — could presumably be employed successfully in these settings.

1.3 Contribution and roadmap

In this paper, we test Kleiner et al.'s hypothesis by conducting a simulation study to compare the BLmnB's performance against that of the bootstrap and the BLB in two well-documented cases of bootstrap inconsistency: (1) Estimating the variability of μ from a $N(\mu, 1)$ distribution where the parameter space for μ is restricted to $[0, \infty)$, and (2) estimating the variability of θ from a $\text{Unif}(0, \theta)$ distribution. We find that while the BLmnB performs comparably to the BLB in the first setting, it is capable of outperforming both the bootstrap and the BLB in the second setting. However, in both settings, the accuracy of the BLmnB is sensitive to the selection of the resample size m.

The remainder of this paper will proceed as follows. In Section 2, we introduce our notational setting and describe the bootstrap, BLB, and BLmnB in detail. We formalize the BLmnB algorithm and identify the steps at which it differs from the BLB. Section 3 establishes the framework for our simulation study. We provide an overview of the two settings mentioned above in which the bootstrap fails, define our choice of estimator quality assessment, and describe our tuning parameter grid search procedure. We present the results of our simulations in Section 4. Finally, we conclude in Section 5 by summarizing our results, reflecting on the limitations of the present study, and proposing several avenues of future research.

2 Notation and algorithms

2.1 Notation

For ease of comparison between the BLmnB and its predecessor, we adopt the notation used by Kleiner et al. (2014), applying some minor modifications as needed and borrowing others from Garnatz (2015). Let $X_1, ..., X_n$ denote an independent and identically distributed random sample of size n drawn from some unknown population P. We denote the empirical distribution corresponding to this sample as $\mathbb{P}_n$, where $(\mathbb{P}_n)_t = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq t)$. Suppose that we are interested in some unknown population parameter $\theta(P)$, and that we estimate $\theta(P)$ via some estimator $\widehat{\theta}_n(\mathbb{P}_n)$. Let $Q_n(P)$ denote the true sampling distribution of this estimator.

We aim to summarize the sampling distribution $Q_n(P)$ in a way that will help us evaluate the quality of the estimator $\widehat{\theta}_n(\mathbb{P}_n)$ — e.g., by computing its variance or the expectation of its bias. Let $\xi(Q_n(P))$ denote some assessment of the quality of $\widehat{\theta}_n(\mathbb{P}_n)$, where the function ξ represents the procedure used to obtain this assessment — e.g., the formation of a confidence interval for $\widehat{\theta}_n(\mathbb{P}_n)$. Our objective is to compute $\xi(Q_n(P))$, but we cannot do so directly since P and $Q_n(P)$ are unknown in practice. Therefore, we must use the random sample $X_1, ..., X_n$ to estimate $\xi(Q_n(P))$. In Sections 2.2, 2.3, and 2.4, we describe three resampling methods designed to accomplish this task.

2.2 The nonparametric bootstrap

Efron's bootstrap (1979) provides a straightforward and accurate computational method for approximating $Q_n(P)$ and estimating $\xi(Q_n(P))$ on the basis of a single sample. Bootstrap estimation of $\xi(Q_n(P))$ can be viewed as a three-step procedure. First, since the true population is unknown, we "plug in" the empirical distribution $\mathbb{P}_n$ for P. In other words, we treat the random sample $X_1, ..., X_n$ as if it were the population, which shifts our focus from approximating $Q_n(P)$ to approximating $Q_n(\mathbb{P}_n)$. This plug-in approximation is justified by the Glivenko-Cantelli theorem (Tucker, 1959), which states that the empirical distribution converges uniformly to the true population distribution — i.e.,

$$\sup_{t\in\mathbb{R}} |(\mathbb{P}_n)_t - P_t| \overset{n\to\infty}{\longrightarrow} 0. \tag{1}$$

Second, we form a Monte Carlo approximation $\mathbb{Q}_n^*(\mathbb{P}_n)$ for $Q_n(\mathbb{P}_n)$ by sampling with replacement from $\mathbb{P}_n$ to generate r resamples of size n, where r is some large integer. We compute a realization of our estimator $\widehat{\theta}_n(\mathbb{P}_n)$ on each of these r resamples and aggregate the r realizations into an empirical sampling distribution $\mathbb{Q}_n^*(\mathbb{P}_n)$. Finally, we apply our estimator quality assessment ξ to this empirical sampling distribution and obtain the bootstrap estimate $\xi(\mathbb{Q}_n^*(\mathbb{P}_n))$ of $\xi(Q_n(P))$.

We summarize these steps in Algorithm 1 below. The brevity of this algorithm reveals that the bootstrap is a very automatic procedure. The only tuning parameter that the user must specify to implement the bootstrap is the number of Monte Carlo iterations r, and the choice of r is somewhat arbitrary as long as it is sufficiently large. However, as discussed in Section 1.2, the bootstrap requires the repeated application of an estimator and estimator quality assessment to this sufficiently large number of resamples, each of which contains approximately $0.632n$ distinct data points from the original sample. As a result, the procedure becomes computationally inefficient for large n.

Algorithm 1 The nonparametric bootstrap

```
Input: Random sample $X_1, ..., X_n$; number of Monte Carlo iterations $r$; estimator quality assessment $\xi$
Output: An estimate of $\xi(Q_n(P))$

For $k \in \{1, ..., r\}$   {
    Sample $n$ iid observations from $\mathbb{P}_n$ with replacement to form the $k$th resample $X^*_{1,k}, ..., X^*_{n,k}$

    Denote the empirical distribution corresponding to the $k$th resample as $\mathbb{P}^*_{n,k}$ , where $(\mathbb{P}^*_{n,k})_t = \frac{1}{n}\sum_{i=1}^{n} I(X^*_{i,k} \leq t)$

    Compute the $k$th estimate $\widehat{\theta}^*_{n,k}(\mathbb{P}^*_{n,k})$
}

Form the empirical sampling distribution $\mathbb{Q}^*_n(\mathbb{P}_n)$, where $(\mathbb{Q}^*_n(\mathbb{P}_n))_t = \frac{1}{r}\sum_{k=1}^{r}(\widehat{\theta}^*_{n,k}(\mathbb{P}^*_{n,k}) \leq t)$

Compute $\xi(\mathbb{Q}^*_n(\mathbb{P}_n))$, the nonparametric bootstrap estimate of $\xi(Q_n(P))$
```

2.3 Bag of little bootstraps (BLB)

The bag of little bootstraps (BLB) (Kleiner et al., 2014) remedies the bootstrap's computational shortcomings in the massive data setting and yields estimator quality assessments that are as or more accurate than those produced by the bootstrap. It achieves this feat by synthesizing the useful characteristics of bootstrapping and subsampling. In short, the BLB involves subsampling the original sample, applying the bootstrap to each subsample, and aggregating the bootstrap results across the subsamples.

More formally, the first step of the BLB algorithm is to generate s subsamples of size $b = n^\gamma$ from the random sample $X_1, ..., X_n$, where $\gamma \in (0, 1)$. We form each subsample by randomly sampling a set of indices $\mathcal{I} = \{i_1, ..., i_b\}$ without replacement from the integers $\{1, ..., n\}$ and designating the set of observations $X_{i_1}, ..., X_{i_b}$ as the subsample. We denote the empirical distribution of the jth subsample as $\mathbb{P}^{(j)}_{n,b}$, where

$$(\mathbb{P}^{(j)}_{n,b})_t = \frac{1}{b}\sum_{i \in \mathcal{I}} I(X_i \leq t). \tag{2}$$

Next, we utilize the plug-in principle described in Section 2.2 and treat the empirical distribution of each subsample as if it were the population. Because of this substitution, we can apply the bootstrap to each subsample. We generate r resamples of size n from the empirical distribution of each subsample, where r is some large number. Note that while each resample is of size n, it contains at most b distinct data points from the original sample. This crucial detail contributes the two major advantages that the BLB holds over subsampling and the bootstrap: (1) Since the resamples have the same nominal size as the original sample, the BLB does not require knowledge of the estimator's convergence rate in order to recover the favorable theoretical properties of the bootstrap, and (2) the BLB only requires repeated computation on samples of size $0.632b << 0.632n$, as each resample contains at most b distinct data points.

As proposed by Kleiner et al. (2014), we use a multinomial random variable to generate the r resamples from each subsample. Specifically, we randomly sample a vector of counts $(n_1, ..., n_b)$ from an n-trial multinomial distribution over b objects, and we form each resulting resample as

Algorithm 2 Bag of little bootstraps (BLB) (Kleiner et al., 2014)

Input: Random sample $X_1, ..., X_n$; number of Monte Carlo iterations r; number of subsamples s; subsample size $b = n^\gamma$, where $\gamma \in (0,1)$; estimator quality assessment ξ

Output: An estimate of $\xi(Q_n(P))$

```
For $j \in \{1, ..., s\}$   {
    Randomly sample a set $\mathcal{I} = \{i_1, ..., i_b\}$ of $b$ indices from $\{1, ..., n\}$ without replacement

    Form the $j$th subsample $X_{i_1}, ..., X_{i_b}$

    Denote the empirical distribution corresponding to the $j$th subsample as $\mathbb{P}^{(j)}_{n,b}$, where $(\mathbb{P}^{(j)}_{n,b})_t = \frac{1}{b}\sum_{i\in\mathcal{I}} I(X_i \leq t)$

    For $k \in \{1, ..., r\}$   {
        Randomly sample a vector of counts $(n_1, ..., n_b) \sim \text{Multinomial}(n, \frac{\mathbf{1}_b}{b})$

        Denote the $k$th resample as $\text{rep}(X_{i_1}, n_1), ..., \text{rep}(X_{i_b}, n_b)$

        Denote the empirical distribution corresponding to the $k$th resample as $\mathbb{P}^*_{n,k}$,
            where $(\mathbb{P}^*_{n,k})_t = \frac{1}{n}\sum_{a=1}^{b} n_a \cdot I(X_{i_a} \leq t)$

        Compute the $k$th estimate $\widehat{\theta}^*_{n,k}(\mathbb{P}^*_{n,k})$
    }

    Form the $j$th empirical sampling distribution $\mathbb{Q}^*_{n,j}(\mathbb{P}^{(j)}_{n,b})$, where $(\mathbb{Q}^*_{n,j}(\mathbb{P}^{(j)}_{n,b}))_t = \frac{1}{r}\sum_{k=1}^{r} I(\widehat{\theta}^*_{n,k}(\mathbb{P}^*_{n,k}) \leq t)$

    Compute the $j$th estimate $\xi^*_j(\mathbb{Q}^*_{n,j}(\mathbb{P}^{(j)}_{n,b}))$ of $\xi(Q_n(P))$
}

Compute $\xi(\mathbb{Q}^*_n(\mathbb{P}_n)) = \frac{1}{s}\sum_{j=1}^{s} \xi(\mathbb{Q}^*_{n,j}(\mathbb{P}^{(j)}_{n,b}))$, the BLB estimate of $\xi(Q_n(P))$
```

the set of n_1 copies of X_{i_1}, n_2 copies of X_{i_2}, etc. For each subsample j, we denote the empirical distribution of the kth resample as $\mathbb{P}^*_{n,k}$, where $k \in \{1, ..., r\}$ and

$$(\mathbb{P}^*_{n,k})_t = \frac{1}{n}\sum_{a=1}^{b} n_a \cdot I(X_{i_a} \leq t), \tag{3}$$

and we compute the kth estimate $\widehat{\theta}^*_{n,k}(\mathbb{P}^*_{n,k})$ on this empirical distribution.

We aggregate the r estimates for each subsample j into the jth empirical sampling distribution $\mathbb{Q}^*_{n,j}(\mathbb{P}^{(j)}_{n,b})$, where

$$(\mathbb{Q}^*_{n,j}(\mathbb{P}^{(j)}_{n,b}))_t = \frac{1}{r}\sum_{k=1}^{r} I(\widehat{\theta}^*_{n,k}(\mathbb{P}^*_{n,k}) \leq t), \tag{4}$$

and we use this empirical sampling distribution to compute the jth estimator quality assessment $\xi^*_j(\mathbb{Q}^*_{n,j}(\mathbb{P}^{(j)}_{n,b}))$. Finally, we take the mean of the s estimator quality assessments to produce a final assessment $\xi(\mathbb{Q}^*_n(\mathbb{P}_n))$, given by

$$\xi(\mathbb{Q}_n^*(\mathbb{P}_n)) = \frac{1}{s}\sum_{j=1}^{s}\xi(\mathbb{Q}_{n,j}^*(\mathbb{P}_{n,b}^{(j)})). \tag{5}$$

The quantity in (5) is the BLB estimate of $\xi(Q_n(P))$. Note that we could also consolidate the s estimator quality assessments using something other than the mean, such as the median or the interquartile mean. Garnatz (2015) evaluates these other aggregation methods in the context of the BLB, but finds that the mean achieves the highest approximation accuracy out of all methods considered. Because of this, and because Kleiner et al. (2014) use the mean in their initial proposal of the BLB, we will also use the mean in the final aggregation step for both the BLB and the BLmnB (see Section 2.4).

Algorithm 2 provides a formal summary of the BLB algorithm. In Algorithm 2, we use $\text{rep}(w, z)$ to denote a list of z copies of some element w, and we use $\mathbf{1}_b$ to denote a b-length vector of ones.

2.4 Bag of little m out of n bootstraps (BLmnB)

While the BLB is expected to efficiently produce accurate results in settings in which the bootstrap succeeds, it is not expected to perform well when the bootstrap fails (Kleiner et al., 2014). However, since the demand for valid assessments of estimator quality does not disappear in these nonideal settings, there is a need to develop new data-driven inferential techniques that are applicable in cases of bootstrap and BLB failure.

Kleiner et al. posit that a small modification of their BLB algorithm — in which resamples of size $m < n$ are generated from each subsample — may perform well even when the bootstrap and BLB fail. The m out of n bootstrap generally restores consistency in cases of bootstrap inconsistency (Samworth, 2003), so replacing the bootstrap with the m out of n bootstrap in the BLB algorithm would presumably have an analogous effect in cases of BLB inconsistency. In this section, we implement Kleiner et al.'s suggested change to the BLB algorithm and formalize the procedure for the new method, which we henceforth refer to as the bag of little m out of n bootstraps (BLmnB).

The BLmnB algorithm (see Algorithm 3) is nearly identical to the BLB algorithm, so we need only highlight the steps at which it differs. First, the BLmnB requires the specification of an additional tuning parameter: the resample size $m = b^\alpha = n^{\alpha\gamma}$, where $\gamma \in (0, 1)$ and $\alpha \in [1, 1/\gamma)$. Note that the restriction of α to $[1, 1/\gamma)$ creates bounds of $[b, n)$ for m — i.e., the resample size is semi-bounded below by the subsample size and strictly bounded above by the original sample size. For comparison, recall that the resample size in the BLB algorithm is automatically set to n.

Like the BLB, the BLmnB first generates s subsamples of size $b = n^\gamma$ from the random sample $X_1, ..., X_n$. In both algorithms, we apply the plug-in principle and treat the empirical distribution of each subsample as if it were the population, and we utilize a multinomial random variable to generate r resamples from each subsample. However, while the BLB draws resamples of size n from each subsample, the BLmnB resamples are of size $m < n$. Specifically, we randomly sample a vector of counts $(m_1, ..., m_b)$ from an m-trial multinomial distribution over b objects, and we form the resulting resample as the set of m_1 copies of X_{i_1}, m_2 copies of X_{i_2}, etc. Therefore, for each subsample j, the empirical distribution of the kth BLmnB resample is denoted as $\mathbb{P}_{n,k}^*$, where $k \in \{1, ..., r\}$ and

Algorithm 3 Bag of little m out of n bootstraps (BLmnB)

Input: Random sample $X_1, ..., X_n$; number of Monte Carlo iterations r; number of subsamples s; subsample size $b = n^\gamma$, where $\gamma \in (0, 1)$; resample size $m = b^\alpha = n^{\alpha\gamma}$, where $\alpha \in [1, 1/\gamma)$; estimator quality assessment ξ

Output: An estimate of $\xi(Q_n(P))$

```
For $j \in \{1, ..., s\}$   {
    Randomly sample a set $\mathcal{I} = \{i_1, ..., i_b\}$ of $b$ indices from $\{1, ..., n\}$ without replacement

    Form the $j$th subsample $X_{i_1}, ..., X_{i_b}$

    Denote the empirical distribution corresponding to the $j$th subsample as $\mathbb{P}^{(j)}_{n,b}$, where $(\mathbb{P}^{(j)}_{n,b})_t = \frac{1}{b}\sum_{i\in\mathcal{I}} I(X_i \le t)$

    For $k \in \{1, ..., r\}$   {
        Randomly sample a vector of counts $(m_1, ..., m_b) \sim \text{Multinomial}(m, \frac{\mathbf{1}_b}{b})$

        Denote the $k$th resample as $\text{rep}(X_{i_1}, m_1), ..., \text{rep}(X_{i_b}, m_b)$

        Denote the empirical distribution corresponding to the $k$th resample as $\mathbb{P}^*_{n,k}$,
            where $(\mathbb{P}^*_{n,k})_t = \frac{1}{m}\sum_{a=1}^{b} m_a \cdot I(X_{i_a} \le t)$

        Compute the $k$th estimate $\widehat{\theta}^*_{n,k}(\mathbb{P}^*_{n,k})$
    }

    Form the $j$th empirical sampling distribution $\mathbb{Q}^*_{n,j}(\mathbb{P}^{(j)}_{n,b})$, where $(\mathbb{Q}^*_{n,j}(\mathbb{P}^{(j)}_{n,b}))_t = \frac{1}{r}\sum_{k=1}^{r} I(\widehat{\theta}^*_{n,k}(\mathbb{P}^*_{n,k}) \le t)$

    Compute the $j$th estimate $\xi^*_j(\mathbb{Q}^*_{n,j}(\mathbb{P}^{(j)}_{n,b}))$ of $\xi(Q_n(P))$
}

Compute $\xi(\mathbb{Q}^*_n(\mathbb{P}_n)) = \frac{1}{s}\sum_{j=1}^{s} \xi(\mathbb{Q}^*_{n,j}(\mathbb{P}^{(j)}_{n,b}))$, the BLB estimate of $\xi(Q_n(P))$
```

$$(\mathbb{P}^*_{n,k})_t = \frac{1}{m}\sum_{a=1}^{b} m_a \cdot I(X_{i_a} \le t). \tag{6}$$

We compute the kth estimate $\widehat{\theta}^*_{n,k}(\mathbb{P}^*_{n,k})$ on this empirical distribution.

The remainder of the BLmnB algorithm is the same as the BLB algorithm. For each subsample, we assemble the r estimates into an empirical sampling distribution, and we compute an estimate of the desired estimator quality assessment on this empirical sampling distribution. We aggregate the resulting s estimator quality assessments into a single estimate by taking the mean, which yields the BLmnB estimate $\xi(\mathbb{Q}^*_n(\mathbb{P}_n))$ of $\xi(Q_n(P))$.

We hypothesize that the BLmnB retains the computational efficiency of the BLB; like the BLB, it involves the repeated computation of estimates on subsets containing only around $0.632b$ distinct data points. More importantly for the context of this paper, we hypothesize that the BLmnB is capable of outperforming both the bootstrap and the BLB in cases of bootstrap failure, conditional on adequate tuning parameter selection. In the following two sections, we formulate and conduct a simulation study to test this second hypothesis by applying the three resampling methods in two settings in which the bootstrap is known to be inconsistent. We elect not to investigate the first hypothesis, although one could feasibly do so by evaluating the runtimes of the three methods in addition to their approximation accuracies.

3 Simulation framework

3.1 Two cases of bootstrap inconsistency

We will assess the potential value of the BLmnB by comparing its sampling distribution approximation accuracy to that of the bootstrap and the BLB in two well-documented settings of bootstrap inconsistency.

3.1.1 Setting 1: $N(\mu, 1)$ where $\mu \in [0, \infty)$

We first consider a boundary problem in which an independent and identically distributed random sample $X_1, ..., X_n$ is drawn from a $N(\mu, 1)$ distribution where the parameter μ is restricted to be nonnegative and the true data-generating μ is equal to zero. The maximum likelihood estimator of μ in this setting is

$$\widehat{\mu}_n = \max\{0, \overline{X}_n\}, \tag{7}$$

the maximum of 0 and the sample mean $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ (Andrews, 2000). Andrews demonstrates that the bootstrap is inconsistent in this setting, as it fails to yield asymptotically correct approximations of the sampling distribution of $\widehat{\mu}_n$.

In our simulation study, we first evaluate the performance of the three resampling methods in a more ideal setting without the restriction on μ — i.e., we draw from a $N(\mu, 1)$ distribution where $-\infty < \mu < \infty$. We refer to this set of simulations as case (a) of setting 1. The maximum likelihood estimator $\widehat{\mu}_n$ in case (a) is the sample mean $\overline{X}_n$, and the bootstrap is consistent since there is no restriction on the parameter space. Hence, this set of simulations is intended to serve as a baseline, as we expect all three methods to perform well under these circumstances. According to Kleiner et al. (2014), the BLB should outperform the bootstrap in case (a), and we expect the BLmnB to do the same since it is a variant of the BLB.

We juxtapose these baseline results with those from the scenario of interest described above, in which we sample from a $N(\mu, 1)$ distribution where $\mu \in [0, \infty)$ and the data-generating μ is equal to zero. We refer to this set of simulations as case (b) of setting 1. In case (b), we expect the bootstrap to perform poorly given its inconsistency, and we expect the BLB to perform relatively poorly as well. We hypothesize that the BLmnB may achieve higher accuracy than the other two methods in this setting.

We reiterate that the data in case (b) are generated from a $N(\mu, 1)$ distribution with a true mean of $\mu = 0$. If μ is small but nonzero, the inconsistency of the bootstrap is less noticeable in practice and the respective accuracies of the three methods lie somewhere between their accuracies in case (a) and case (b). Specifically, when the true mean is small and nonzero, the BLB and the bootstrap perform similarly to case (a) for smaller values of γ (i.e., smaller subsamples) and similarly to case (b) for larger values of γ (i.e., larger subsamples). See Appendix A for simulation results corresponding to an alternative version of case (b) in which the data-generating μ is 0.05.

3.1.2 Setting 2: Unif$(0, \theta)$

The second setting we consider involves estimating the variability of the endpoint θ of a continuous uniform distribution. Suppose that we draw an independent and identically distributed random sample $X_1, ..., X_n$ from a Unif$(0, \theta)$ distribution. The maximum likelihood estimator of θ is

$$\widehat{\theta}_n = \max\{X_1, ..., X_n\}, \tag{8}$$

the maximum order statistic of the sample (Hogg et al., 2019). Bickel and Freedman (1981) show that the bootstrap is inconsistent in this setting. Hence, this is another scenario where the bootstrap fails when the parameter of interest lies on the boundary of the parameter space. We expect the bootstrap and the BLB to perform relatively poorly in this setting due to the inconsistency of the bootstrap, and we hypothesize that the BLmnB may overcome this issue and achieve higher approximation accuracy than the other two methods.

Unlike in setting 1, there is no baseline scenario to consider in this setting. Thus, we analyze only one case in our simulation study for setting 2. We generate data from a $\text{Unif}(0, \theta)$ distribution where the true θ is set arbitrarily to one.

3.2 Simulation procedure

Recall that the output of each of the bootstrap, BLB, and BLmnB algorithms is an estimate $\xi(\mathbb{Q}_n^*(\mathbb{P}_n))$ of some estimator quality assessment $\xi(Q_n(P))$. Therefore, in order to evaluate these three methods, we must select an assessment procedure ξ and use simulated data to compute the "ground truth" assessment $\xi(Q_n(P))$. We can then compare the estimates produced by the bootstrap, BLB, and BLmnB to the ground truth to determine the accuracy with which each method approximates the sampling distribution $Q_n(P)$.

We adopt the majority of our simulation methodology from Kleiner et al. (2014), although we apply some modifications since our two settings of interest are both boundary problems. We define $\xi(Q_n(P))$ as a 95% percentile confidence interval for the parameter of interest, and we use the width of this confidence interval as the benchmark against which we will compare our bootstrap, BLB, and BLmnB estimates. Note that the exact form of $\xi(Q_n(P))$ varies slightly for the different settings described in Section 3.1 — $\xi(Q_n(P))$ is a two-sided interval for μ in case (a) of setting 1, a lower one-sided interval for μ in case (b) of setting 1, and an upper one-sided interval for θ in setting 2.

To compute the ground truth $\xi(Q_n(P))$ in each setting, we generate 10,000 independent and identically distributed random samples of size $n = 20,000$ from the true distribution P. We compute a realization of the maximum likelihood estimator on each sample, assemble these 10,000 estimates into a high-fidelity approximation of the sampling distribution $Q_n(P)$, and form the relevant confidence interval $\xi(Q_n(P))$. We then compute the width of this true confidence interval.

Next, we generate a new independent and identically distributed random sample of size $n = 20,000$ from the true distribution P. We run each of the three methods on this sample using some prespecified combination of tuning parameters (see Section 3.3), and we obtain an estimate $\xi(\mathbb{Q}_n^*(\mathbb{P}_n))$ for each method. To assess the accuracy of each estimated confidence interval $\xi(\mathbb{Q}_n^*(\mathbb{P}_n))$, we compute its width and compare this to the ground truth confidence interval width. Specifically, we compute the relative error

$$\frac{|w^* - w|}{w}, \tag{9}$$

where w^* denotes the width of the estimated confidence interval $\xi(\mathbb{Q}_n^*(\mathbb{P}_n))$ and w denotes the width of the ground truth confidence interval $\xi(Q_n(P))$.

We repeat the process from the preceding paragraph for nine more independent and identically distributed random samples of size $n = 20,000$. Thus, for each of the bootstrap, BLB, and BLmnB, we compute the relative error ten times for each combination of tuning parameters. We take the average of these ten relative errors to obtain a single value that summarizes the approximation accuracy of each method for a particular combination of tuning parameters. These average relative errors are the quantities reported in Figures 1, 2, and 3.

3.3 Tuning parameter selection

The approximation accuracies of the bootstrap and BLB are affected by the tuning parameters of their respective algorithms (Garnatz, 2015; Kleiner et al., 2014), and it is logical to assume that the same is true for the BLmnB. In the following table, we summarize the tuning parameters that affect the performance of these three methods:

Bootstrap	**BLB**	**BLmnB**
Number of Monte Carlo iterations r	Number of Monte Carlo iterations r	Number of Monte Carlo iterations r
	Number of subsamples s	Number of subsamples s
	Subsample size $b = n^\gamma$	Subsample size $b = n^\gamma$
		Resample size $m = b^\alpha = n^{\alpha\gamma}$

We examine the effects of these tuning parameters in our simulation study by repeating the procedure from Section 3.2 for several feasible values of r, s, γ (which determines the subsample size b), and α (which determines the BLmnB resample size m). For r, s, and γ, we consider a range of values based on the simulations conducted by Kleiner et al. (2014):

- $r \in \{25, 50, 100, 250, 500\}$
- $s \in \{10, 20, 40, 100, 200\}$
- $\gamma \in \{0.5, 0.7, 0.9\}$

We have defined α to lie in the half-open interval $[1, 1/\gamma)$ to ensure that the BLmnB resample size falls between b and n, but there exist no additional suggestions in the literature regarding a range of reasonable values for α. Bickel and Sakov (2008) studied the choice of the resample size m in context of the m out of n bootstrap, but this work is not directly applicable to the BLmnB since it does not account for the preliminary subsampling stage. As such, we present a novel grid search strategy that specifies a vector of evenly spaced of values of α.

We assume that we want to conduct our simulations for a range of BLmnB resample sizes that (1) has a lower bound some $100x\%$ (where $x \in [0, 1]$) of the way between the subsample size $b = n^\gamma$ and the original sample size n, and (2) has an upper bound that is close to, but smaller than, the original sample size n. In other words, we want $m = n^{\alpha\gamma}$ to have a lower bound of $n^\gamma + x(n - n^\gamma)$ and an upper bound that is arbitrarily close to n (e.g., $n^{0.999}$). Under these assumptions, it can be shown that the desired lower bound α_{LB} and upper bound α_{UB} for the tuning parameter α are

$$\alpha_{\text{LB}} = \frac{1}{\gamma} \cdot \frac{\log(n^\gamma + x(n - n^\gamma))}{\log(n)} \qquad \text{and} \qquad \alpha_{\text{UB}} = \frac{0.999}{\gamma}, \tag{10}$$

respectively.

In our simulation study, we set x equal to 0.25 for cases (a) and (b) of setting 1 and equal to zero for setting 2. For a given γ, we grid search over a sequence of ten evenly spaced values of α between α_{LB} and α_{UB}, which yields a corresponding sequence of ten values of m. However, for the sake of parsimony, we report our results in Section 4 and Appendix A for a smaller selection of values of m.

4 Simulation results

4.1 Setting 1

4.1.1 Setting 1, case (a)

Figure 1 reports the results of our simulations for case (a) of setting 1, in which we generate data from a $N(\mu, 1)$ distribution where $-\infty < \mu < \infty$ and the true μ is equal to zero. For the BLB and BLmnB, we report in Figure 1 only the results for $\gamma = 0.7$; see Appendix A for additional results corresponding to $\gamma = 0.5$ and $\gamma = 0.9$.

In this baseline case, we find that the bootstrap, BLB, and BLmnB all achieve a low average relative error for nearly all tuning parameter combinations. This is consistent with our hypothesis from Section 3.1.1. Since there is no restriction on the parameter space in case (a), it is not surprising that the three methods perform well overall.

While they share in this overall accuracy, there are some notable differences in performance between the three methods. For each resample size r, the BLB achieves a lower average relative error than the bootstrap. This is consistent with the results from Kleiner et al. (2014), which similarly suggest that the BLB outperforms the bootstrap in this type of ideal setting with no parameter space restrictions.

The approximation accuracy of the BLmnB is comparable to that of the BLB (and, hence, better than that of the bootstrap) for sufficiently large values of the resample size m. This "sufficiently large" threshold appears to be at a value of approximately $m = 15,000$, or $0.75n$. The performance of the BLmnB is quite sensitive to the choice of m; while the BLmnB performs comparably to the BLB when m is greater than approximately 15,000, it performs worse than both the BLB and the bootstrap when m is smaller than this threshold.

While changes to s and γ seem to have a minimal effect on average relative error for the BLB and BLmnB, there appears to be a meaningful relationship between the number of Monte Carlo iterations r and the average relative error for all three methods. One curious result for the BLmnB that resurfaces in case (b) of setting 1 and in setting 2 is that, for smaller values of m, the average relative error tends to increase as r increases. This is unexpected, as an increase in the number of Monte Carlo iterations generally leads to a decrease in approximation error for bootstrap-based methods. Indeed, this is what we observe for the bootstrap and the BLB in case (a), as well as for the BLmnB with larger values of m. It appears that the BLmnB results converge to the BLB results as m approaches n, which is rather intuitive. However, it is not clear why the expected relationship between r and the average relative error is reversed for smaller values of m. This is one potential starting point for future research on the BLmnB.

4.1.2 Setting 1, case (b)

Figure 2 displays the results of our simulations for case (b) of setting 1, in which we generate data from a $N(\mu, 1)$ distribution where μ is restricted to be nonnegative and the true μ is equal to zero. In Figure 2, we again focus only on the results for $\gamma = 0.7$ for the BLB and BLmnB; see Appendix A for additional results corresponding to $\gamma = 0.5$ and $\gamma = 0.9$.

We find that the bootstrap, BLB, and BLmnB all perform worse in case (b) than they did in case (a). This is not surprising since the bootstrap is no longer consistent in case (b) due to the restric-

Figure 1: Simulation results for setting 1, case (a)

Bootstrap

r	Average relative error
25	0.17
50	0.18
100	0.07
250	0.07
500	0.04

BLB

$\gamma = 0.7$		s				
		10	20	40	100	200
r	25	0.11	0.12	0.13	0.14	0.13
	50	0.09	0.09	0.08	0.08	0.08
	100	0.05	0.05	0.04	0.04	0.04
	250	0.03	0.03	0.01	0.02	0.02
	500	0.02	0.01	0.01	0.01	0.01

BLmnB

$\gamma = 0.7$		s				
		10	20	40	100	200
m = 11,446	r = 25	0.13	0.13	0.14	0.15	0.14
	50	0.21	0.20	0.22	0.21	0.22
	100	0.25	0.27	0.26	0.26	0.26
	250	0.29	0.30	0.30	0.29	0.30
	500	0.31	0.31	0.31	0.31	0.31
m = 13,127	25	0.08	0.06	0.09	0.08	0.07
	50	0.13	0.14	0.14	0.13	0.14
	100	0.19	0.18	0.18	0.18	0.18
	250	0.22	0.21	0.21	0.21	0.21
	500	0.23	0.23	0.22	0.22	0.22
m = 15,056	25	0.04	0.02	0.02	0.01	0.01
	50	0.04	0.07	0.06	0.06	0.07
	100	0.11	0.09	0.10	0.10	0.10
	250	0.13	0.13	0.13	0.13	0.13
	500	0.15	0.14	0.14	0.14	0.14
m = 17,267	25	0.06	0.04	0.07	0.07	0.07
	50	0.04	0.02	0.02	0.01	0.01
	100	0.03	0.03	0.04	0.03	0.03
	250	0.06	0.06	0.05	0.06	0.06
	500	0.06	0.06	0.06	0.06	0.07
m = 19,803	25	0.14	0.14	0.14	0.14	0.13
	50	0.06	0.08	0.08	0.08	0.07
	100	0.03	0.04	0.04	0.04	0.04
	250	0.02	0.02	0.01	0.02	0.01
	500	0.01	0.01	0.01	0.01	0.00

Figure 2: Simulation results for setting 1, case (b)

Bootstrap

r	Average relative error
25	0.46
50	0.39
100	0.42
250	0.38
500	0.40

BLB

$\gamma = 0.7$		s				
		10	20	40	100	200
r	25	0.28	0.16	0.16	0.11	0.16
	50	0.28	0.27	0.16	0.13	0.09
	100	0.45	0.19	0.15	0.22	0.13
	250	0.39	0.18	0.22	0.21	0.20
	500	0.37	0.24	0.29	0.24	0.17

BLmnB

$\gamma = 0.7$		s				
		10	20	40	100	200
m = 11,446	r = 25	0.29	0.37	0.24	0.23	0.21
	50	0.47	0.34	0.41	0.37	0.33
	100	0.42	0.52	0.42	0.40	0.32
	250	0.50	0.45	0.37	0.46	0.49
	500	0.81	0.53	0.50	0.54	0.53
m = 13,127	25	0.50	0.21	0.20	0.15	0.17
	50	0.30	0.17	0.20	0.25	0.28
	100	0.45	0.28	0.30	0.27	0.27
	250	0.29	0.47	0.34	0.36	0.37
	500	0.49	0.35	0.45	0.41	0.39
m = 15,056	25	0.38	0.30	0.21	0.14	0.16
	50	0.22	0.20	0.24	0.21	0.18
	100	0.25	0.38	0.25	0.26	0.25
	250	0.35	0.52	0.32	0.24	0.27
	500	0.42	0.31	0.35	0.31	0.28
m = 17,267	25	0.19	0.28	0.18	0.10	0.11
	50	0.25	0.20	0.13	0.14	0.10
	100	0.18	0.29	0.22	0.19	0.14
	250	0.28	0.31	0.27	0.23	0.20
	500	0.32	0.41	0.25	0.28	0.28
m = 19,803	25	0.32	0.22	0.16	0.17	0.13
	50	0.40	0.16	0.22	0.10	0.12
	100	0.25	0.27	0.19	0.19	0.12
	250	0.25	0.24	0.25	0.16	0.14
	500	0.34	0.15	0.25	0.25	0.18

tion on μ. The BLB achieves a lower average relative error than the bootstrap for nearly all tuning parameter combinations, and it performs particularly well relative to the bootstrap for larger values of s. Thus, it appears that the BLB's advantage over the bootstrap extends to this nonideal setting.

Just like in case (a), the BLmnB performs comparably to the BLB (and, therefore, better than the bootstrap) provided that the resample size m is sufficiently large. In fact, for each value of γ, the BLmnB results appears to converge to those of the BLB as m approaches n. While the BLmnB achieves a lower average relative error than the BLB for certain tuning parameter combinations, it does not appear to do so systematically. Hence, we do not encounter strong evidence in support of our hypothesis from Section 3.1.1 that the BLmnB outperforms the BLB in case (b).

There does not appear to be a meaningful relationship between average relative error and r for the bootstrap and the BLB. For the BLmnB, the average relative error tends to increase as r increases for smaller values of m, but this relationship dissipates for larger values of m. As discussed in Section 4.1.1, we generally expect Monte Carlo approximation error to decrease as the number of iterations increases. It is not clear why this relationship does not hold in our results for case (b).

There also does not appear to be a clear relationship between average relative error and the subsample size $b = n^\gamma$ for either the BLB or the BLmnB — the average relative error does not seem to systematically increase or decrease as γ increases. However, for both of these methods, the average relative error appears to vary more across the various combinations of r and s for smaller values of γ (i.e., smaller subsamples) than for larger values of γ (i.e., larger subsamples).

Unlike r and γ, the number of subsamples s seems to have an effect on average relative error for the BLB and BLmnB. For both of these methods, the average relative error tends to decrease as s increases. However, this trend is more noticeable for $\gamma = 0.5$ and $\gamma = 0.7$ than for $\gamma = 0.9$.

4.2 Setting 2

Figure 3 and Appendix A present the results of our simulations for setting 2, in which we generate data from a Unif$(0, \theta)$ distribution where the true θ is equal to one. For the BLB and BLmnB, we report in Figure 3 and Appendix A only the results for $\gamma = 0.9$, as we encounter a practical issue when running these two methods using values of γ smaller than 0.9.

This practical issue arises because the estimator of interest in setting 2 is an extreme order statistic. Suppose that we run the BLB in setting 2 using a value of $\gamma = 0.7$. Since $n = 20,000$, this means that each subsample contains $b = 20,000^{0.7} \approx 1,025$ distinct data points from the original sample. Therefore, when we generate BLB resamples of size $n = 20,000$ from each subsample of size $b \approx 1,025$, each element of the subsample is almost certain to appear in each resample at least once (and likely around 19 or 20 times, on average). This implies that the maximum order statistic of each resample — i.e., each realization of our estimator $\widehat{\theta}_n$ — is almost certain to be the maximum value of the subsample. As a result, the sampling distribution of $\widehat{\theta}_n$ for each subsample will have all of its mass concentrated at the maximum of the subsample, and the estimated confidence interval for each subsample will have a width of zero. In turn, the BLB estimate $\xi(\mathbb{Q}_n^*(\mathbb{P}_n))$ for the confidence interval will also have a width of zero, so the corresponding relative error will be equal to $\frac{|0-w|}{w} = 1$. In other words, the BLB is unable to produce a meaningful approximation of the sampling distribution when the estimator of interest is an extreme order statistic. The BLmnB essentially encounters the same problem. It is possible to overcome the issue with the BLmnB by

Figure 3: Simulation results for setting 2

Bootstrap

r	Average relative error
25	0.62
50	0.58
100	0.63
250	0.51
500	0.51

BLB

$\gamma = 0.9$		s				
		10	20	40	100	200
r	25	0.69	0.68	0.70	0.63	0.66
	50	0.59	0.57	0.56	0.58	0.59
	100	0.44	0.44	0.53	0.53	0.52
	250	0.44	0.34	0.37	0.44	0.41
	500	0.50	0.25	0.36	0.30	0.32

BLmnB

$\gamma = 0.9$		s				
		10	20	40	100	200
m = 9,237	r = 25	0.44	0.32	0.51	0.36	0.35
	50	0.55	0.52	0.51	0.46	0.49
	100	0.49	0.46	0.58	0.50	0.55
	250	0.76	0.76	0.58	0.61	0.63
	500	0.60	0.64	0.61	0.63	0.66
m = 10,301	25	0.26	0.36	0.16	0.20	0.20
	50	0.40	0.27	0.34	0.22	0.24
	100	0.36	0.33	0.36	0.31	0.26
	250	0.45	0.44	0.41	0.41	0.39
	500	0.53	0.56	0.58	0.53	0.48
m = 11,486	25	0.33	0.24	0.20	0.20	0.16
	50	0.23	0.19	0.23	0.15	0.14
	100	0.15	0.22	0.24	0.24	0.15
	250	0.25	0.21	0.18	0.15	0.17
	500	0.28	0.23	0.16	0.19	0.15
m = 12,808	25	0.24	0.20	0.25	0.16	0.21
	50	0.19	0.27	0.21	0.17	0.17
	100	0.28	0.22	0.20	0.25	0.21
	250	0.30	0.28	0.27	0.27	0.24
	500	0.36	0.30	0.25	0.26	0.28
m = 14,282	25	0.31	0.38	0.29	0.29	0.27
	50	0.31	0.24	0.25	0.23	0.22
	100	0.25	0.31	0.25	0.24	0.24
	250	0.33	0.37	0.27	0.27	0.30
	500	0.31	0.30	0.27	0.25	0.24
m = 15,926	25	0.40	0.38	0.35	0.36	0.42
	50	0.46	0.36	0.25	0.32	0.34
	100	0.30	0.28	0.30	0.30	0.27
	250	0.31	0.35	0.30	0.27	0.30
	500	0.30	0.32	0.26	0.25	0.29

setting the resample size m to be sufficiently small relative to n, but this decision may not be wise since it often comes at the expense of approximation accuracy.

Hence, for the BLB and BLmnB, we focus only on the results for $\gamma = 0.9$. We observe that the BLB achieves a lower average relative error than the bootstrap for sufficiently large values of r. This "sufficiently large" threshold appears to be at a value of approximately $r = 100$. Overall, both the bootstrap and the BLB perform relatively poorly, which is consistent with our hypothesis from Section 3.1.2.

Also consistent with our hypothesis from Section 3.1.2 is the fact that the BLmnB systematically outperforms both the bootstrap and the BLB for certain values of m. While the BLmnB performs comparably to or worse than the other two methods for relatively small and relatively large values of m, there exists a range of intermediate values of m for which the BLmnB achieves a lower average relative error for all tuning parameter combinations. These findings suggest that there exists an optimal value of m somewhere in this intermediate range (which, in this case, appears to be between approximately $11,000$ and $13,000$). As such, the choice of m is crucial to the success of the BLmnB in the $\text{Unif}(0, \theta)$ setting.

While the number of subsamples s does not appear to have a clear effect on the average relative error for either the BLB or the BLmnB, the number of Monte Carlo iterations r seems to have a meaningful impact for all three methods. For the bootstrap and the BLB, the average relative error tends to decrease as r increases, which is consistent with our expectations for Monte Carlo approximation accuracy. For the BLmnB, we observe the same strange relationship between r and the average relative error that we observed in setting 1: The error tends to increase as r increases for smaller values of m, but it tends to decrease as r increases for larger values of m. As such, the BLmnB results seem to converge to the BLB results as m approaches n.

5 Discussion

5.1 Summary of results

The bag of little bootstraps (BLB) is an accurate and computationally efficient method of estimator quality assessment for massive data sets, but it is not expected to perform well in settings in which the bootstrap is inconsistent (Kleiner et al., 2014). In this paper, we formalized and evaluated the bag of little m out of n bootstrap (BLmnB), a modification of the BLB that aims to extend the method's applicability to cases of bootstrap inconsistency without sacrificing its accuracy or computational benefits.

Our simulation results indicate that while the BLmnB is capable of achieving higher sampling distribution approximation accuracy than the bootstrap and the BLB in at least some cases of bootstrap failure, its success is not guaranteed in all such cases. On the one hand, we found that the BLmnB systematically outperforms the bootstrap and the BLB for several values of the resample size m when the parameter of interest is the endpoint θ of a $\text{Unif}(0, \theta)$ distribution. However, when the parameter of interest is the mean μ of a $N(\mu, 1)$ distribution where μ is restricted to be nonnegative, the BLmnB performs no better than the BLB. In both settings, we found that the success of the BLmnB is sensitive to the choice of m.

5.2 Limitations and suggestions for future research

While this paper makes a valuable contribution to the resampling literature by formalizing the BLmnB algorithm and demonstrating the potential value of the method via simulation, our work is not without its limitations. In particular, we identify four ways in which our presentation of the BLmnB is limited in scope.

First, we did not develop the necessary theory to show that the BLmnB shares the favorable theoretical properties — i.e., consistency and higher order correctness — of the bootstrap and the BLB. We mentioned in Section 2.4 that substituting the m out of n bootstrap for the bootstrap in the BLB algorithm would presumably recover consistency in cases of bootstrap inconsistency, but we did not demonstrate this argument formally. One could presumably do so by making the appropriate modifications to the theoretical results presented by Kleiner et al. (2014). It is important that this theoretical foundation is established before further BLmnB simulation studies are conducted.

Second, we considered in our simulations only one choice of the estimator quality assessment $\xi(Q_n(P))$. For ease of comparison with the original BLB proposal, we adopted Kleiner et al.'s (2014) definition of $\xi(Q_n(P))$ — we defined $\xi(Q_n(P))$ as a 95% percentile confidence interval, and we used the width of this confidence interval to compare the bootstrap, BLB, and BLmnB. While it is plausible that the relative performance of the three methods would be similar for alternative quality assessments such as the variance or bias of the estimator, it would be worthwhile to verify this claim through additional simulations.

Third, we evaluated the approximation accuracy of the three methods in only two settings of bootstrap inconsistency. Our results indicate that the relative performance of the BLmnB is better in the $\text{Unif}(0, \theta)$ setting than in the restricted $N(\mu, 1)$ setting, which implies that each case of bootstrap inconsistency is somewhat unique. In other words, our conclusions from setting 1 and setting 2 may not be generalizable to other cases of bootstrap inconsistency. Andrews (2000) provides an overview of several other counterexamples to the consistency of the bootstrap; it would be informative to assess the performance of the bootstrap, BLB, and BLmnB in these settings.

Finally, while we hypothesized in Section 2.4 that the BLmnB retains the computational efficiency of the BLB, we did not test this hypothesis in our simulations. Like the BLB, the BLmnB involves the repeated computation of estimates on subsets of approximately $0.632b << 0.632n$ distinct data points, so it is reasonable to assume that the BLmnB shares — or even enhances — the BLB's computational advantage over the bootstrap. One could evaluate this assumption by repeating our simulation procedure and recording the respective runtimes of the bootstrap, BLB, and BLmnB for different tuning parameter combinations.

The most productive avenue of future research on the BLmnB would involve expanding the scope of this paper to address the four limitations discussed above. However, there are two other aspects of our simulation results that similarly require further investigation. First, it remains unclear why the BLmnB is capable of performing better than the bootstrap and the BLB in the $\text{Unif}(0, \theta)$ setting, but not in the restricted $N(\mu, 1)$ setting. We infer that this discrepancy is related to the fact that the two settings present different types of boundary problems with different maximum likelihood estimators, but one would need to investigate the theory underlying these cases of bootstrap inconsistency in order to validate this claim.

Second, while we noted that there appears to be an optimal resample size m for the BLmnB algorithm in the $\text{Unif}(0, \theta)$ setting, it is not obvious how one might determine this optimal m in practice. It would be useful to develop an iterative tuning parameter selection procedure for the BLmnB, similar to those proposed by Kleiner et al. (2014) and Bickel and Sakov (2008). Such a procedure would greatly improve the automatic capacity of the BLmnB and make it an even more appealing method of estimator quality assessment in the massive data setting.

References

[1] Andrews, D. W. K. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, *68*(2), 399–405.

[2] Bickel, P. J., Götze, F., & van Zwet, W. R. (2012). Resampling fewer than n observations: Gains, losses, and remedies for losses. In *Selected Works of Willem van Zwet* (pp. 267–297). Springer.

[3] Bickel, Peter J., & Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics*, *9*(6), 1196–1217.

[4] Bickel, Peter J., & Sakov, A. (2008). On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statistica Sinica*, *18*(3), 967–985.

[5] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*(1), 1–26.

[6] Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall.

[7] Garnatz, C. (2015). *Trusting the black box: Confidence with bag of little bootstraps*. Pomona College.

[8] Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer.

[9] Hogg, R. V., McKean, J. W., & Craig, A. T. (2019). *Introduction to Mathematical Statistics* (8th ed.). Pearson.

[10] Kleiner, A., Talwalkar, A., Sarkar, P., & Jordan, M. I. (2014). A scalable bootstrap for massive data. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *76*(4), 795–816.

[11] Politis, D. N., Romano, J. P., & Wolf, M. (1999). *Subsampling*. Springer.

[12] Putter, H., & van Zwet, W. R. (1996). Resampling: consistency of substitution estimators. *Annals of Statistics*, *24*(6), 2297–2318.

[13] Samworth, R. (2003). A note on methods of restoring consistency to the bootstrap. *Biometrika*, *90*(4), 985–990.

[14] Tucker, H. G. (1959). A generalization of the Glivenko-Cantelli theorem. *The Annals of Mathematical Statistics*, *30*(3), 828–830.

[15] Tukey, J. (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, *29*, 614.

Appendix A: Additional results

Setting 1, case (a): $N(\mu, 1)$ where $-\infty < \mu < \infty$, true $\mu = 0$

Bootstrap

r	Average relative error
25	0.17
50	0.18
100	0.07
250	0.07
500	0.04

BLB

$\gamma = 0.5$		s				
		10	20	40	100	200
r	25	0.13	0.14	0.14	0.14	0.14
	50	0.09	0.09	0.09	0.09	0.08
	100	0.05	0.05	0.05	0.05	0.05
	250	0.02	0.03	0.03	0.02	0.03
	500	0.02	0.03	0.02	0.01	0.01

$\gamma = 0.7$		s				
		10	20	40	100	200
r	25	0.11	0.12	0.13	0.14	0.13
	50	0.09	0.09	0.08	0.08	0.08
	100	0.05	0.05	0.04	0.04	0.04
	250	0.03	0.03	0.01	0.02	0.02
	500	0.02	0.01	0.01	0.01	0.01

$\gamma = 0.9$		s				
		10	20	40	100	200
r	25	0.12	0.14	0.13	0.13	0.13
	50	0.09	0.08	0.08	0.08	0.08
	100	0.04	0.04	0.05	0.04	0.04
	250	0.02	0.02	0.02	0.02	0.02
	500	0.01	0.01	0.01	0.01	0.01

BLmnB

$\gamma = 0.5$		s				
		10	20	40	100	200
m = 10,842	r = 25	0.19	0.19	0.19	0.18	0.17
	50	0.25	0.25	0.23	0.24	0.25
	100	0.28	0.30	0.30	0.30	0.29
	250	0.32	0.32	0.32	0.32	0.32
	500	0.35	0.34	0.35	0.34	0.33
m = 12,604	25	0.10	0.08	0.08	0.09	0.08
	50	0.14	0.13	0.16	0.15	0.15
	100	0.19	0.20	0.20	0.20	0.20
	250	0.25	0.23	0.23	0.23	0.23
	500	0.24	0.24	0.24	0.24	0.24
m = 14,653	25	0.05	0.02	0.02	0.02	0.02
	50	0.08	0.06	0.08	0.07	0.07
	100	0.10	0.11	0.10	0.11	0.11
	250	0.14	0.14	0.13	0.14	0.14
	500	0.16	0.16	0.15	0.14	0.15
m = 17,034	25	0.06	0.08	0.08	0.06	0.07
	50	0.02	0.02	0.02	0.01	0.01
	100	0.03	0.05	0.04	0.03	0.03
	250	0.05	0.07	0.06	0.06	0.06
	500	0.07	0.08	0.08	0.06	0.06
m = 19,803	25	0.12	0.12	0.14	0.14	0.13
	50	0.08	0.09	0.08	0.08	0.09
	100	0.06	0.04	0.04	0.04	0.04
	250	0.04	0.02	0.02	0.02	0.02
	500	0.02	0.02	0.01	0.01	0.01

BLmnB (cont.)

$\gamma = 0.7$				s		
		10	20	40	100	200
m = 11,446	r = 25	0.13	0.13	0.14	0.15	0.14
	50	0.21	0.20	0.22	0.21	0.22
	100	0.25	0.27	0.26	0.26	0.26
	250	0.29	0.30	0.30	0.29	0.30
	500	0.31	0.31	0.31	0.31	0.31
m = 13,127	25	0.08	0.06	0.09	0.08	0.07
	50	0.13	0.14	0.14	0.13	0.14
	100	0.19	0.18	0.18	0.18	0.18
	250	0.22	0.21	0.21	0.21	0.21
	500	0.23	0.23	0.22	0.22	0.22
m = 15,056	25	0.04	0.02	0.02	0.01	0.01
	50	0.04	0.07	0.06	0.06	0.07
	100	0.11	0.09	0.10	0.10	0.10
	250	0.13	0.13	0.13	0.13	0.13
	500	0.15	0.14	0.14	0.14	0.14
m = 17,267	25	0.06	0.04	0.07	0.07	0.07
	50	0.04	0.02	0.02	0.01	0.01
	100	0.03	0.03	0.04	0.03	0.03
	250	0.06	0.06	0.05	0.06	0.06
	500	0.06	0.06	0.06	0.06	0.07
m = 19,803	25	0.14	0.14	0.14	0.14	0.13
	50	0.06	0.08	0.08	0.08	0.07
	100	0.03	0.04	0.04	0.04	0.04
	250	0.02	0.02	0.01	0.02	0.01
	500	0.01	0.01	0.01	0.01	0.00

BLmnB (cont.)

$\gamma = 0.9$				s		
		10	20	40	100	200
m = 14,982	r = 25	0.06	0.02	0.02	0.01	0.01
	50	0.07	0.08	0.06	0.06	0.07
	100	0.11	0.10	0.11	0.11	0.11
	250	0.14	0.14	0.13	0.14	0.13
	500	0.14	0.14	0.14	0.15	0.14
m = 16,065	25	0.04	0.04	0.05	0.03	0.04
	50	0.05	0.02	0.02	0.03	0.02
	100	0.07	0.05	0.07	0.06	0.07
	250	0.09	0.10	0.10	0.09	0.10
	500	0.10	0.10	0.10	0.10	0.11
m = 17,225	25	0.06	0.06	0.07	0.07	0.06
	50	0.03	0.02	0.02	0.01	0.01
	100	0.04	0.03	0.03	0.03	0.03
	250	0.05	0.06	0.06	0.06	0.06
	500	0.07	0.07	0.06	0.07	0.07
m = 18,469	25	0.10	0.11	0.10	0.10	0.10
	50	0.05	0.05	0.04	0.05	0.05
	100	0.02	0.02	0.02	0.01	0.01
	250	0.02	0.02	0.02	0.02	0.02
	500	0.03	0.03	0.03	0.03	0.03
m = 19,803	25	0.12	0.11	0.12	0.12	0.13
	50	0.07	0.07	0.08	0.07	0.07
	100	0.03	0.04	0.04	0.04	0.04
	250	0.01	0.02	0.01	0.02	0.01
	500	0.01	0.01	0.01	0.00	0.00

Setting 1, case (b): $N(\mu, 1)$ where $\mu \in [0, \infty)$, true $\mu = 0$

Bootstrap

r	Average relative error
25	0.46
50	0.39
100	0.42
250	0.38
500	0.40

BLB

$\gamma = 0.5$				s		
		10	20	40	100	200
r	25	0.36	0.21	0.23	0.10	0.08
	50	0.37	0.21	0.12	0.11	0.11
	100	0.29	0.25	0.23	0.28	0.20
	250	0.60	0.52	0.40	0.26	0.24
	500	0.56	0.38	0.31	0.36	0.39

$\gamma = 0.7$				s		
		10	20	40	100	200
r	25	0.28	0.16	0.16	0.11	0.16
	50	0.28	0.27	0.16	0.13	0.09
	100	0.45	0.19	0.15	0.22	0.13
	250	0.39	0.18	0.22	0.21	0.20
	500	0.37	0.24	0.29	0.24	0.17

$\gamma = 0.9$				s		
		10	20	40	100	200
r	25	0.44	0.37	0.37	0.32	0.34
	50	0.31	0.37	0.30	0.30	0.27
	100	0.29	0.28	0.33	0.31	0.30
	250	0.22	0.34	0.26	0.30	0.30
	500	0.39	0.33	0.33	0.30	0.29

BLmnB

$\gamma = 0.5$				s		
		10	20	40	100	200
m = 10,842	r = 25	0.47	0.60	0.42	0.34	0.39
	50	0.48	0.68	0.52	0.53	0.48
	100	0.79	0.62	0.50	0.63	0.58
	250	0.61	0.93	0.64	0.64	0.67
	500	0.98	0.68	0.88	0.80	0.80
m = 12,604	25	0.28	0.38	0.28	0.25	0.22
	50	0.48	0.38	0.42	0.33	0.37
	100	0.37	0.50	0.51	0.49	0.59
	250	0.81	0.55	0.52	0.59	0.55
	500	0.79	0.61	0.74	0.59	0.74
m = 14,653	25	0.29	0.17	0.25	0.17	0.18
	50	0.42	0.33	0.26	0.27	0.30
	100	0.47	0.44	0.40	0.35	0.34
	250	0.45	0.53	0.61	0.48	0.43
	500	0.68	0.52	0.47	0.52	0.58
m = 17,034	25	0.24	0.20	0.08	0.15	0.08
	50	0.35	0.37	0.27	0.21	0.20
	100	0.39	0.47	0.35	0.29	0.26
	250	0.72	0.35	0.45	0.36	0.33
	500	0.70	0.43	0.47	0.47	0.47
m = 19,803	25	0.20	0.12	0.11	0.07	0.08
	50	0.38	0.17	0.22	0.11	0.13
	100	0.33	0.29	0.27	0.26	0.21
	250	0.25	0.24	0.30	0.18	0.24
	500	0.37	0.29	0.31	0.38	0.37

BLmnB (cont.)

$\gamma = 0.7$		s				
		10	20	40	100	200
m = 11,446	r = 25	0.29	0.37	0.24	0.23	0.21
	50	0.47	0.34	0.41	0.37	0.33
	100	0.42	0.52	0.42	0.40	0.32
	250	0.50	0.45	0.37	0.46	0.49
	500	0.81	0.53	0.50	0.54	0.53
m = 13,127	25	0.50	0.21	0.20	0.15	0.17
	50	0.30	0.17	0.20	0.25	0.28
	100	0.45	0.28	0.30	0.27	0.27
	250	0.29	0.47	0.34	0.36	0.37
	500	0.49	0.35	0.45	0.41	0.39
m = 15,056	25	0.38	0.30	0.21	0.14	0.16
	50	0.22	0.20	0.24	0.21	0.18
	100	0.25	0.38	0.25	0.26	0.25
	250	0.35	0.52	0.32	0.24	0.27
	500	0.42	0.31	0.35	0.31	0.28
m = 17,267	25	0.19	0.28	0.18	0.10	0.11
	50	0.25	0.20	0.13	0.14	0.10
	100	0.18	0.29	0.22	0.19	0.14
	250	0.28	0.31	0.27	0.23	0.20
	500	0.32	0.41	0.25	0.28	0.28
m = 19,803	25	0.32	0.22	0.16	0.17	0.13
	50	0.40	0.16	0.22	0.10	0.12
	100	0.25	0.27	0.19	0.19	0.12
	250	0.25	0.24	0.25	0.16	0.14
	500	0.34	0.15	0.25	0.25	0.18

BLmnB (cont.)

$\gamma = 0.9$		s				
		10	20	40	100	200
m = 14,982	r = 25	0.28	0.26	0.30	0.30	0.27
	50	0.34	0.35	0.33	0.32	0.26
	100	0.28	0.27	0.29	0.31	0.30
	250	0.40	0.29	0.23	0.29	0.25
	500	0.24	0.27	0.31	0.29	0.29
m = 16,065	25	0.17	0.33	0.30	0.30	0.29
	50	0.31	0.29	0.32	0.29	0.29
	100	0.34	0.30	0.31	0.26	0.28
	250	0.36	0.35	0.33	0.29	0.30
	500	0.26	0.29	0.35	0.27	0.28
m = 17,225	25	0.35	0.33	0.30	0.32	0.28
	50	0.48	0.31	0.32	0.31	0.32
	100	0.32	0.25	0.29	0.31	0.28
	250	0.23	0.39	0.34	0.26	0.28
	500	0.31	0.38	0.23	0.28	0.30
m = 18,469	25	0.38	0.29	0.31	0.31	0.30
	50	0.33	0.33	0.27	0.29	0.31
	100	0.26	0.31	0.26	0.27	0.29
	250	0.36	0.34	0.28	0.27	0.31
	500	0.44	0.25	0.35	0.30	0.27
m = 19,803	25	0.27	0.34	0.33	0.30	0.29
	50	0.25	0.37	0.31	0.28	0.31
	100	0.30	0.32	0.32	0.33	0.30
	250	0.38	0.34	0.30	0.33	0.34
	500	0.26	0.38	0.30	0.30	0.29

Setting 1, alternative version of case (b): $N(\mu, 1)$ where $\mu \in [0, \infty)$, true $\mu = 0.05$

Bootstrap

r	Average relative error
25	0.17
50	0.18
100	0.07
250	0.07
500	0.04

BLB

$\gamma = 0.5$		s				
		10	20	40	100	200
r	25	0.35	0.43	0.37	0.40	0.39
	50	0.34	0.36	0.34	0.36	0.35
	100	0.25	0.30	0.32	0.32	0.31
	250	0.28	0.25	0.27	0.33	0.30
	500	0.21	0.34	0.31	0.31	0.28

$\gamma = 0.7$		s				
		10	20	40	100	200
r	25	0.17	0.17	0.19	0.20	0.20
	50	0.15	0.16	0.16	0.14	0.14
	100	0.10	0.08	0.11	0.10	0.11
	250	0.05	0.10	0.08	0.08	0.08
	500	0.10	0.08	0.08	0.08	0.08

$\gamma = 0.9$		s				
		10	20	40	100	200
r	25	0.12	0.14	0.13	0.13	0.13
	50	0.09	0.08	0.08	0.08	0.08
	100	0.04	0.04	0.05	0.04	0.04
	250	0.02	0.02	0.02	0.02	0.02
	500	0.01	0.01	0.01	0.01	0.01

BLmnB

$\gamma = 0.5$		s				
		10	20	40	100	200
m = 10,842	r = 25	0.12	0.06	0.18	0.17	0.16
	50	0.23	0.08	0.11	0.10	0.13
	100	0.13	0.10	0.08	0.08	0.09
	250	0.11	0.10	0.10	0.10	0.07
	500	0.16	0.13	0.06	0.06	0.06
m = 12,604	25	0.26	0.19	0.23	0.21	0.24
	50	0.21	0.18	0.19	0.18	0.19
	100	0.16	0.16	0.14	0.16	0.13
	250	0.12	0.14	0.13	0.13	0.14
	500	0.10	0.13	0.11	0.14	0.12
m = 14,653	25	0.31	0.34	0.28	0.29	0.29
	50	0.29	0.26	0.27	0.26	0.24
	100	0.21	0.23	0.20	0.24	0.23
	250	0.17	0.18	0.15	0.18	0.20
	500	0.19	0.21	0.17	0.18	0.18
m = 17,034	25	0.32	0.36	0.34	0.34	0.35
	50	0.34	0.26	0.30	0.31	0.30
	100	0.30	0.24	0.25	0.26	0.26
	250	0.13	0.22	0.23	0.25	0.26
	500	0.22	0.25	0.22	0.23	0.23
m = 19,803	25	0.42	0.39	0.40	0.40	0.38
	50	0.33	0.32	0.34	0.36	0.35
	100	0.28	0.29	0.31	0.31	0.32
	250	0.28	0.34	0.29	0.33	0.30
	500	0.34	0.36	0.29	0.29	0.29

BLmnB (cont.)

$\gamma = 0.7$				s		
		10	20	40	100	200
m = 11,446	r = 25	0.07	0.06	0.07	0.06	0.06
	50	0.10	0.10	0.12	0.12	0.13
	100	0.16	0.19	0.17	0.17	0.16
	250	0.21	0.21	0.18	0.20	0.20
	500	0.25	0.23	0.23	0.22	0.22
m = 13,127	25	0.05	0.06	0.03	0.02	0.02
	50	0.07	0.06	0.05	0.07	0.05
	100	0.07	0.08	0.09	0.09	0.10
	250	0.16	0.14	0.12	0.12	0.12
	500	0.18	0.13	0.13	0.14	0.12
m = 15,056	25	0.06	0.09	0.08	0.08	0.07
	50	0.04	0.05	0.05	0.04	0.02
	100	0.05	0.06	0.04	0.02	0.03
	250	0.08	0.05	0.06	0.06	0.06
	500	0.11	0.08	0.06	0.06	0.06
m = 17,267	25	0.12	0.10	0.14	0.13	0.14
	50	0.11	0.08	0.07	0.07	0.08
	100	0.07	0.06	0.07	0.05	0.04
	250	0.06	0.03	0.03	0.03	0.02
	500	0.06	0.04	0.02	0.02	0.02
m = 19,803	25	0.21	0.21	0.21	0.21	0.19
	50	0.11	0.13	0.13	0.14	0.13
	100	0.11	0.10	0.10	0.10	0.11
	250	0.12	0.09	0.08	0.09	0.08
	500	0.13	0.07	0.08	0.08	0.07

BLmnB (cont.)

$\gamma = 0.9$				s		
		10	20	40	100	200
m = 14,982	r = 25	0.06	0.02	0.02	0.01	0.01
	50	0.07	0.08	0.06	0.06	0.07
	100	0.11	0.10	0.11	0.11	0.11
	250	0.13	0.14	0.13	0.14	0.13
	500	0.14	0.14	0.14	0.15	0.14
m = 16,065	25	0.04	0.04	0.05	0.03	0.04
	50	0.05	0.02	0.02	0.03	0.02
	100	0.07	0.05	0.07	0.06	0.07
	250	0.09	0.10	0.10	0.09	0.10
	500	0.10	0.10	0.10	0.10	0.11
m = 17,225	25	0.06	0.06	0.07	0.07	0.06
	50	0.03	0.02	0.02	0.01	0.01
	100	0.04	0.03	0.03	0.03	0.03
	250	0.05	0.06	0.06	0.06	0.06
	500	0.07	0.07	0.06	0.07	0.07
m = 18,469	25	0.10	0.11	0.10	0.10	0.10
	50	0.05	0.05	0.04	0.05	0.05
	100	0.02	0.02	0.02	0.01	0.01
	250	0.02	0.02	0.02	0.02	0.02
	500	0.03	0.03	0.03	0.03	0.03
m = 19,803	25	0.12	0.11	0.12	0.12	0.13
	50	0.07	0.07	0.08	0.07	0.07
	100	0.03	0.04	0.04	0.04	0.04
	250	0.01	0.02	0.01	0.02	0.01
	500	0.01	0.01	0.01	0.00	0.00

Setting 2: Unif$(0, \theta)$ where true $\theta = 1$

Bootstrap

r	Average relative error
25	0.62
50	0.58
100	0.63
250	0.51
500	0.51

BLB

$\gamma = 0.9$		s				
		10	20	40	100	200
r	25	0.69	0.68	0.70	0.63	0.66
	50	0.59	0.57	0.56	0.58	0.59
	100	0.44	0.44	0.53	0.53	0.52
	250	0.44	0.34	0.37	0.44	0.41
	500	0.50	0.25	0.36	0.30	0.32

BLmnB

$\gamma = 0.9$		s				
		10	20	40	100	200
m = 7,429	r = 25	0.88	0.90	0.88	0.87	0.87
	50	1.09	1.01	0.97	0.89	0.95
	100	1.05	0.94	0.88	1.01	0.97
	250	1.03	0.93	1.04	1.05	1.03
	500	1.07	1.11	1.00	1.09	1.04
m = 8,284	r = 25	0.63	0.55	0.54	0.58	0.58
	50	0.87	0.57	0.69	0.64	0.66
	100	0.78	0.91	0.86	0.77	0.74
	250	0.67	0.73	0.67	0.73	0.72
	500	0.68	0.78	0.56	0.71	0.67
m = 9,237	r = 25	0.44	0.32	0.51	0.36	0.35
	50	0.55	0.52	0.51	0.46	0.49
	100	0.49	0.46	0.58	0.50	0.55
	250	0.76	0.76	0.58	0.61	0.63
	500	0.60	0.64	0.61	0.63	0.66
m = 10,301	r = 25	0.26	0.36	0.16	0.20	0.20
	50	0.40	0.27	0.34	0.22	0.24
	100	0.36	0.33	0.36	0.31	0.26
	250	0.45	0.44	0.41	0.41	0.39
	500	0.53	0.56	0.58	0.53	0.48
m = 11,486	r = 25	0.33	0.24	0.20	0.20	0.16
	50	0.23	0.19	0.23	0.15	0.14
	100	0.15	0.22	0.24	0.24	0.15
	250	0.25	0.21	0.18	0.15	0.17
	500	0.28	0.23	0.16	0.19	0.15

BLmnB (cont.)

$\gamma = 0.9$ (cont.)				s		
		10	20	40	100	200
m = 12,808	r = 25	0.24	0.20	0.25	0.16	0.21
	50	0.19	0.27	0.21	0.17	0.17
	100	0.28	0.22	0.20	0.25	0.21
	250	0.30	0.28	0.27	0.27	0.24
	500	0.36	0.30	0.25	0.26	0.28
m = 14,282	r = 25	0.31	0.38	0.29	0.29	0.27
	50	0.31	0.24	0.25	0.23	0.22
	100	0.25	0.31	0.25	0.24	0.24
	250	0.33	0.37	0.27	0.27	0.30
	500	0.31	0.30	0.27	0.25	0.24
m = 15,926	r = 25	0.40	0.38	0.35	0.36	0.42
	50	0.46	0.36	0.25	0.32	0.34
	100	0.30	0.28	0.30	0.30	0.27
	250	0.31	0.35	0.30	0.27	0.30
	500	0.30	0.32	0.26	0.25	0.29
m = 17,759	r = 25	0.48	0.57	0.51	0.52	0.50
	50	0.52	0.43	0.42	0.43	0.45
	100	0.36	0.37	0.30	0.31	0.37
	250	0.29	0.31	0.35	0.26	0.30
	500	0.32	0.35	0.28	0.29	0.26
m = 19,803	r = 25	0.69	0.63	0.62	0.65	0.65
	50	0.63	0.62	0.54	0.59	0.55
	100	0.53	0.52	0.48	0.46	0.46
	250	0.44	0.37	0.38	0.36	0.38
	500	0.34	0.36	0.31	0.30	0.30