FAST AND ROBUST BOOTSTRAP IN ANALYSING LARGE MULTIVARIATE DATASETS

Shahab Basiri, Esa Ollila & Visa Koivunen

Aalto University, Dept. of Signal Processing and Acoustics, P.O.Box 13000, FI-00076 Aalto, Finland

ABSTRACT

In this paper we address the problem of performing statistical inference for large scale data sets. The volume and dimensionality of the data may be so high that it cannot be processed or stored in a single node. We propose a scalable, statistically robust and computationally efficient bootstrap method compatible with distributed processing and storage systems. Bootstrapping is performed on multiple smaller distinct subsets of data similarly to the bag of little bootstrap method (BLB) [1]. For each bootstrap replica drawn from distinct data subsets, a computationally efficient fixed-point estimation equation is solved. The proposed bootstrap method facilitates using highly robust statistical methods in analyzing large scale data sets. Significant savings in computation is achieved since the method does not require recomputing the estimator for each bootstrap sample but it is done analytically using a smart approximation. Simulation examples demonstrate the usefulness and validity of the method for bootstrap analysis of large data sets.

Index Terms— bootstrap, bag of little bootstraps, fast and robust bootstrap, big data, robust estimation, distributed computation.

1. INTRODUCTION

Recent advances in digital technology have led to a proliferation of big data sets. Examples include climate data, social networking, smart phone and health data, etc. Processing and storage of massive data sets becomes possible through parallel and distributed architectures. However, the volume of the data has grown to an extent that cannot be effectively handled by traditional statistical analysis and inferential methods. Performing statistical inference on massive data sets using distributed and parallel platforms require fundamental changes in statistical methodology. Even estimation of a parameter of interest based on the entire massive data set can be prohibitively expensive and assigning estimates of uncertainty (error bars, confidence intervals, etc.) to statistical estimates [3, 4]. However, for at least two obvious reasons the method is computationally impractical for analysis of modern high volume and high-dimensional data sets: First, the size of each bootstrap sample is the same as the original big data set (with about 63% of data points appearing at least once in each sample), thus leading to processing and storage problems even in advanced computing systems. Second, (re)computation of value of the estimator for each massive bootstrapped data set is not feasible even for estimators with moderate level of computational complexity. Variants such as subsampling [5] and the m out of n bootstrap [6] were proposed to reduce the computational cost of bootstrap by computation of the point estimates on smaller sub-samples of the original data set. Implementation of such methods is even more problematic as the output is sensitive to the size of the sub-samples m. In addition extra analytical effort is needed in order to re-scale the output to the right size.

The bag of little bootstraps (BLB) [1] modifies the Efron’s bootstrap to make it applicable for massive data sets. In BLB method the massive data is subdivided randomly into disjoint subsets (i.e., so called subsample modules or bags). This allows the massive data sets to be stored in distributed fashion. Moreover subsample modules can be processed in parallel using distributed computing architectures. The BLB samples are constructed by assigning random weights from multinomial distribution to the data points of a disjoint subsample. Although in BLB the problem of handling and processing massive bootstrap samples is alleviated, yet (re)computation of the estimates for a large number of bootstrap samples is prohibitively expensive. Hence, the method may be impractical for many commonly used modern estimators that typically have a high complexity. Such estimators often require solving demanding optimization problems numerically.

In this paper we address the problem of bootstrapping massive data sets by introducing a low complexity and robust bootstrap method. The new method possesses similar scalability property as BLB scheme with significantly lower computational complexity. Low complexity is achieved by utilizing the Fast and Robust Bootstrap (FRB) technique [8, 9, 2] which avoids (re)computation of fixed-point equations for each bootstrap sample via a smart approximation. Although
the FRB method possesses a lower complexity in comparison with the conventional bootstrap, the original FRB is incompatible with distributed processing and storage platforms and not suitable for bootstrap analysis of massive data sets. Our proposed bootstrap method is scalable and compatible with distributed computing architectures and storage systems, robust to outliers and provides more accurate results in a much faster rate than the original BLB method.

The paper is organized as follows. In Section 2, the BLB and FRB methods are reviewed. The new bootstrap scheme is proposed in Section 3, followed by implementation of the method for MM-estimator of regression [10]. Section 4 provides simulation examples and Section 5 concludes.

2. RELATED BOOTSTRAP METHODS

Here we describe the main idea of the BLB [1] and FRB [2] methods as they are closely related to our proposed bootstrap scheme. The pros and cons of each method are discussed as well.

2.1. Bag of Little Bootstraps

Let \( X = (x_1, \cdots, x_n) \in \mathbb{R}^{m \times n} \) be an \( m \) dimensional observed data set of size \( n \). The volume and dimensionality of data may be so high that it cannot be processed or stored in a single node. Consider \( \hat{\theta}_n \in \mathbb{R}^m \) as an estimator of a parameter of interest \( \theta \in \mathbb{R}^m \) based on \( X \). Computation of estimate of uncertainty \( \hat{\xi} \) (e.g., confidence intervals, standard deviation, etc.) for \( \theta \) is of great interest as for large data sets confidence intervals are often more informative than point estimates.

The bag of little bootstraps (BLB) [1] is a scalable bootstrap scheme that draws disjoint subsamples (which form "bags" or "modules") of smaller size \( b = \lfloor n^\gamma \rfloor \in [0.6, 0.9] \) by randomly resampling without replacement from columns of \( X \). For example if \( n = 10000000 \) and \( \gamma = 0.6 \), then \( b = 15849 \). Furthermore, the computation for each subsample can be done in parallel by different computing nodes. For each subsample module, bootstrap samples, \( X^\ast \), are generated by assigning a random weight vector \( n^\ast = (n_1^\ast, \ldots, n_b^\ast) \) from \( \text{Multinomial}(n, (1/b)1_b) \) to data points of the subsample, where the weights sum to \( n \). Hence, each bootstrap sample contains at most \( b \) distinct data points. Thus the BLB approach produces the bootstrap replicas with reduced effort in comparison to conventional bootstrap [3]. Nevertheless, (re)computing the value of estimator for each bootstrap sample for example thousands of times is still computationally impractical even for estimators of moderate level of complexity. This includes a wide range of modern estimators that are solutions to optimization problems such as maximum likelihood methods or highly robust estimators of linear regression. The desired estimate of uncertainty \( \hat{\xi}^\ast \) is computed based on the population \( \hat{\theta}_n \) within each subsample module and the final estimate is obtained by averaging \( \hat{\xi}^\ast \)'s over the modules.

2.2. Fast and Robust Bootstrap

The fast and robust bootstrap method [2, 8, 9] is computationally efficient and robust to outliers in comparison with Efron’s bootstrap. It is applicable for estimators \( \hat{\theta}_n \in \mathbb{R}^m \) that can be expressed as a solution to a system of smooth fixed-point (FP) equations:

\[
\hat{\theta}_n = Q(\hat{\theta}_n; X),
\]

where \( Q : \mathbb{R}^m \to \mathbb{R}^m \). The bootstrap replicated estimator \( \hat{\theta}^\ast_n \) then solves

\[
\hat{\theta}^\ast_n = Q(\hat{\theta}^\ast_n; X^\ast),
\]

where the function \( Q \) is same as in (1) but now dependent on the bootstrap sample \( X^\ast \). Then, instead of computing \( \hat{\theta}^\ast_n \) from (2), we compute:

\[
\hat{\theta}^\ast_{Rn} = Q(\hat{\theta}_{Rn}; X^\ast),
\]

where the notation \( \hat{\theta}^\ast_{Rn} \) denotes a one-step improvement of the FP iteration (2) with initial value \( \hat{\theta}_{Rn} \) based on bootstrap sample \( X^\ast \). In conventional bootstrap, one uses the distribution of \( \hat{\theta}^\ast_n \) to estimate the sampling distribution of \( \theta_n \). Since the distribution of the one-step estimator \( \hat{\theta}^\ast_{Rn} \) does not accurately reflect the sampling variability of \( \theta \), but typically underestimates it, a linear correction needs to be applied as follows:

\[
\hat{\theta}_{Rn} = \hat{\theta}_n + \left[ I - \nabla Q(\hat{\theta}_n; X) \right]^{-1} (\hat{\theta}^\ast_{Rn} - \hat{\theta}_n),
\]

where \( \nabla Q(\cdot) \in \mathbb{R}^{m \times m} \) is the matrix of partial derivatives w.r.t. \( \theta_n \). Then under sufficient regularity conditions, \( \hat{\theta}_{Rn} \) will be estimating the limiting distribution of \( \theta_n \). In most applications, \( \hat{\theta}^\ast_{Rn} \) is not only significantly faster to compute than \( \hat{\theta}^\ast_n \), but numerically more stable and robust as well. However, FRB is not scalable or compatible with distributed storage and processing systems. Hence, it is not suited for bootstrap analysis of massive data sets. The method has been applied to many complex fixed-point estimators such as FastICA estimator [7], PCA and highly robust estimators of linear regression [2].

3. FAST AND ROBUST BOOTSTRAP FOR BIG DATA

In this section we propose a new bootstrap method that combines the desirable properties of the BLB and FRB methods. The method can be applied to any estimator representable as smooth FP equations. The developed Bag of Little Fast and Robust Bootstraps (BLFRB) method is suitable for big data analysis because of its scalability and low computational complexity. Recall that the main computational burden of the BLB scheme is in recomputation of estimating equation (2)
for each bootstrap sample \( X^* \). Such computational complexity can be drastically reduced by using the FRB replications \( \hat{\theta}_{R^*} \) as in (4). Assuming that the computation of \( \hat{\theta}_n \) based on the original big data set \( X \) is not possible (i.e., since \( X \) might be stored in a distributed manner), we compute the FRB replications \( \hat{\theta}_{R^*} \) locally for bootstrap samples within each bag. The bootstrap samples of BLB can be equivalently constructed by resampling with replacement from \( X = (X; [b/n]_1 \ldots b) \), which is a data set of size \( n \) consists of \( b \) distinct data points each having equal multiplicity of occurrence \([b/n]\). Therefore, in local computation of (4), instead of \( \hat{\theta}_n \), one may compute \( \hat{\theta}_{b} \) based on \( X \). The proposed BLFRB procedure is given in detail in Algorithm 1. The steps of the algorithm are illustrated in Fig. 1, where \( X^{(k)} \), \( k = 1, \ldots, s \) denotes the disjoint subsamples and \( X^{(k)} \) corresponds to the \( j \)th bootstrap replica generated from the distinct subsample \( k \).

While the BLFRB procedure inherits the scalability of BLB, it is radically faster to compute, since the replication \( \hat{\theta}_{R^*} \) can be computed in closed-form with small number of distinct data points. Low complexity of the BLFRB scheme allows for fast and scalable computation of confidence intervals for commonly used modern fixed-point estimators.

### 3.1. BLFRB for MM-estimator of linear regression

Here we present a practical example formulation of the method, where the proposed BLFRB method is used for linear regression. Robust MM-estimator that lends itself to fixed point estimation equations is employed for bootstrap replications. Let \( X = \{ (y_1, z_1^\top), \ldots, (y_n, z_n^\top) \} \), \( z_i \in \mathbb{R}^p \), be a sample of independent random vectors that follow the linear model: \( y_i = z_i^\top \theta + \sigma e_i \), with unknown parameter vector \( \theta \in \mathbb{R}^p \). Noise terms \( e_i \)‘s are i.i.d. random variables from a symmetric distribution with unit scale.

Highly robust MM-estimators [10] are based on two loss functions \( \rho_0 : \mathbb{R} \to \mathbb{R}^+ \) and \( \rho_1 : \mathbb{R} \to \mathbb{R}^+ \) which determine the breakdown point and efficiency of the estimator, respectively. The \( \theta_{\rho_0} \) and \( \theta_{\rho_1} \) functions are symmetric, twice continuously differentiable with \( \rho(0) = 0 \), strictly increasing on \([0, c]\) and constant on \([c, \infty)\) for some constant \( c \). The MM-estimate of \( \hat{\theta}_n \) satisfies

\[
\frac{1}{n} \sum_{i=1}^{n} \rho'(y_i - \hat{z}_i^\top \hat{\theta}_n) \frac{y_i}{\hat{\sigma}_n} = 0
\]

where \( \hat{\sigma}_n \) is a scale S-estimate [11] obtained from minimization of the M-scale \( \hat{\sigma}_n(\theta) \) defined as

\[
\frac{1}{n} \sum_{i=1}^{n} \rho_0 \left( \frac{y_i - \hat{z}_i^\top \theta}{\hat{\sigma}_n(\theta)} \right) = d,
\]

where \( d = \rho_0(\infty)/2 \) is a constant. For future reference let \( \tilde{\theta}_n \) be the argument that minimizes \( \hat{\sigma}_n(\theta) \),

\[
\tilde{\theta}_n = \arg \min_{\theta \in \mathbb{R}^p} \hat{\sigma}_n(\theta).
\]

The Tukey’s loss function:

\[
\rho_c(u) = \begin{cases} \frac{(u^2 - c^2)^2}{2c^2}, & \text{for } |u| \leq c, \\ \frac{c^2}{2} - \frac{c^2}{2}, & \text{for } |u| > c. \end{cases}
\]

is widely used as the \( \rho \) functions of the MM-estimator, where subscript \( c \) represents different tunings of the function. For instance an MM-estimator with efficiency \( O = 95\% \) and breakdown point \( BP = 50\% \) (i.e. for Gaussian errors) is achievable by tuning \( \rho_c(u) \) into \( c_0 = 1.547 \) and \( c_1 = 4.685 \) for \( \rho_0 \) and \( \rho_1 \) respectively (see [12, p.142, tab.19]).

In order to apply the BLFRB method to MM-estimator, (5) and (6) need to be presented in form of FP equations scalable to number of distinct data points in the data. The FP estimating equations are given by [2, eq. 14 and 15] as:

\[
\hat{\theta}_n = \left( \sum_{i=1}^{n} \omega_i z_i z_i^\top + \frac{1}{n} \sum_{i=1}^{n} \omega_i y_i \right)^{-1} \sum_{i=1}^{n} \omega_i z_i y_i,
\]

\[
\hat{\sigma}_n = \sum_{i=1}^{n} \frac{1}{n} \frac{\omega_i}{\hat{\sigma}_n} (\hat{y}_i - \hat{z}_i^\top \hat{\theta}_n),
\]

where \( r_i = y_i - \hat{z}_i^\top \hat{\theta}_n \), \( \tilde{r}_i = y_i - \tilde{z}_i^\top \tilde{\theta}_n \), \( \omega_i = \rho'_1(r_i/\hat{\sigma}_n)/r_i \) and \( v_i = \frac{\omega_i}{\hat{\sigma}_n} \rho_0(\tilde{r}_i/\hat{\sigma}_n)/\tilde{r}_i \). The corresponding scalable one-step MM-estimates \( \hat{\theta}_{n^*} \) and \( \hat{\sigma}_{n^*} \) are obtained by modifying [2, eq. 17 and 18] as follows. Let \( X^* = (X; n^*) \)

---

**Algorithm 1: The BLFRB procedure**

1. Draw \( s \) subsamples (which form “bags” or “modules”) \( \hat{X} = (\hat{x}_1 \cdots \hat{x}_b) \) of smaller size \( b = \{ n^2 \gamma \} \in [0.6, 0.9] \) by randomly sampling without replacement from columns of \( X \);

2. for each subsample \( \hat{X} \) do:
   - Generate \( r \) bootstrap samples by resampling as follows: Bootstrap sample \( \hat{X}^* = (\hat{X}; n^*) \) is formed by assigning a random weight vector \( n^* = (n_1^*, \ldots, n_b^*) \) from \( \text{Multinomial}(n, (1/b)1_b) \) to columns of \( \hat{X} \);

3. Find the initial estimate \( \hat{\theta}_b \) for each distinct subsample \( \hat{X} \) only once and for each bootstrap sample \( \hat{X}^* \) compute \( \hat{\theta}_{R^*} \) from equation (4) using \( \hat{\theta}_b \);

4. Compute the desired estimate of uncertainty \( \hat{\xi}_n \) based on the population of \( r \) FRB replicated values \( \hat{\theta}_{R^*} \);

5. Average the computed values of the estimate of uncertainty over the subsamples, i.e.,

\[
\hat{\xi}_n = \frac{1}{s} \sum_{k=1}^{s} \hat{\xi}^{(k)}.
\]
The steps of the BLFRB procedure (Algorithm 1) are depicted. Disjoint subsamples of significantly smaller size $b$ are drawn from the original Big Data set $X$. The initial estimate $\hat{\theta}_b$ is obtained by solving fixed-point estimating equation only once for each subsample $X$. Within each module, the FRB replicas $\hat{\theta}_n^{R*}$ are computed for each bootstrap sample $X^*$ using the initial estimate $\hat{\theta}_b$. The final estimate of uncertainty $\hat{\xi}$ is obtained by averaging the results of distinct subsample modules.

The initial estimate $\hat{\theta}_b$ is obtained by solving fixed-point estimating equation only once for each subsample $X$. Within each module, the FRB replicas $\hat{\theta}_n^{R*}$ are computed for each bootstrap sample $X^*$ using the initial estimate $\hat{\theta}_b$. The final estimate of uncertainty $\hat{\xi}$ is obtained by averaging the results of distinct subsample modules.

**4. NUMERICAL EXAMPLES**

In this section the performance of the BLFRB method is assessed by simulation studies. We also show the results obtained with the original BLB method for comparison purposes. We generate $n = 50000$ observations from the linear model $y_i = z_i^T \theta + \sigma_0 e_i$, ($i = 1, \ldots, n$), where the explaining variables $z_i$ are generated from $p$-variate normal distribution $N_p(0, I_p)$ with $p = 50$, $p$-dimensional parameter vector $\theta = 1_p$, noise terms are i.i.d. from the standard normal distribution and noise variance is $\sigma_0^2 = 10$. Let $\hat{\theta}_n$ be the finite sample estimate of $\theta$ based on the simulated data set $X = \{(y_1, z_1^T), \ldots, (y_n, z_n^T)\}$. The step 5 of the procedure for the $l$th element of $\hat{\theta}_n$, is as follows:

$$\hat{\xi}_l = \frac{SD(\hat{\theta}_{l,n}^{(k)})}{\left(1 - \frac{1}{r} \sum_{j=1}^{r} (\hat{\theta}_{l,n}^{(kj)} - \bar{\theta}_{\xi,l,n}^{*})^2 \right)^{1/2}},$$

where $\hat{\theta}_{l,n}^{(k)}$ denotes the $l$th element of $\hat{\theta}_n$, $\bar{\theta}_{\xi,l,n}^{*} = \frac{1}{r} \sum_{j=1}^{r} \hat{\theta}_{l,n}^{(kj)}$, and $\xi$ is the step 5 of the procedure for the $l$th element of $\hat{\theta}_n$. The step 5 of the procedure for the $l$th element of $\hat{\theta}_n$ is obtained by:

$$\hat{\xi}_l^* = \frac{SD(\hat{\theta}_{l,n}^{(k)})}{\frac{1}{8} \sum_{h=1}^{8} SD(\hat{\theta}_{l,n}^{(k_h)})}, \quad l = 1, \ldots, p.$$

The performance of the BLB and BLFRB are assessed by computing a relative error defined as:

$$\varepsilon = \frac{|SD(\hat{\theta}_n) - SD_0(\hat{\theta}_n)|}{SD_0(\hat{\theta}_n)},$$

where $SD(\hat{\theta}_n) = \frac{1}{p} \sum_{l=1}^{p} SD(\hat{\theta}_{l,n})$ and $SD_0(\hat{\theta}_n) = \sigma_0/\sqrt{nO}$ is (approximation) of the average standard deviation of $\hat{\theta}_n$ based on the asymptotic covariance matrix [10].
Fig. 2. Relative errors of the BLB (dashed line) and BLFRB (solid line) methods w.r.t. the number of bootstrap samples \( r \) are illustrated. BLFRB achieves significantly lower relative error level than BLB.

Fig. 3. Relative errors of the BLB and BLFRB methods illustrating severe lack of robustness of BLB whereas BLFRB performs highly reliably regardless of the outliers (i.e., \( \mathcal{O} \) is 0.95 for the MM-estimator and 1 for the LS-estimator). The bootstrap setup is as follows; Number of disjoint subsamples is \( s = 25 \), size of each subsample is \( b = \lfloor n^\gamma \rfloor = 1946 \) with \( \gamma = 0.7 \), maximum number of bootstrap samples in each subsample module is \( r_{\text{max}} = 100 \).

We start from \( r = 2 \) and continually add a new set of bootstrap samples (while \( r < r_{\text{max}} \)) to subsample modules. The convergence of relative errors w.r.t. the number of bootstrap samples \( r \) are illustrated in Fig.2. Note that BLFRB achieves significantly lower relative error level.

To study the robustness of the methods, we introduce outliers by multiplying 1% of the original data points by 10. This can be interpreted as misplacement of the decimal points. While such level of outlier contamination ruins the performance of the original BLB method, the propose BLFRB is able to maintain the same level of accuracy. This lack of robustness of the BLB method is illustrated in Fig.3.

Next, let us make an intuitive comparison between computational complexity of the BLB and BLFRB methods by using the MM-estimator in both methods for computation of \( \hat{\theta}_n \) via an identical computing system. The computed errors and the cumulative processing time are stored after each iteration (i.e., adding new set of bootstrap samples to the bags). Fig.4, reports relative errors w.r.t. the required processing time of each iteration of the algorithms. The BLFRB is remarkably faster as in this method the estimating equations need not be solved for each bootstrap sample.

5. CONCLUSION

A new robust, scalable and low complexity bootstrap method is introduced with the aim of finding parameter estimates and confidence measures for very large scale data sets. While the proposed BLFRB method is fully scalable and compatible with distributed computing systems, it is remarkable faster and statistically more robust in comparison to the original BLB method. The statistical properties including convergence and robustness are established in [13].

6. REFERENCES


