Aalto University
School of Electrical
Engineering

Robust, scalable and fast bootstrap method for analyzing large scale data

IEEE SPS Distinguished Lecturer program

Visa Koivunen

Aalto University, School of Electrical Engineering
joint work with Shahab Basiri & Esa Ollila

Fall 2015

Ongoing research topics

- Array signal processing: multiantenna comms
- Wireless communications: flexible spectrum use, wireless localization
- Statistical inference and optimization for smart grids, cyber-physical systems.
- Emerging radar technologies: distributed (MIMO) radar systems, agile/cognitive radars, co-existence of radar and wireless comms systems
- Statistical signal processing theory and methods.

Large Scale Data Analysis

- We live in an era of data deluge
- The lack of scalability of the conventional signal processing (SP) and machine learning techniques and the complexity of data form a bottleneck in the search of relevant information.
- Data may be so BIG that it is not possible to store and process all the data in the same unit.
- We leverage the rich field of statistical signal processing to extract relevant infomation from high-volume and high-dimensional data

Data Analysis problem at hand

- Let $\mathbf{X} = (\mathbf{x}_1 \cdots \mathbf{x}_n) \in \mathbb{R}^{d \times n}$ be a large volume and high dimensional data set that can only be processed and stored via parallel and distributed architectures.
- Let $\hat{\theta}_n$ be an estimator of a parameter of interest $\theta \in \mathbb{R}^d$ based on the observed big data $\mathbf{X}$.
- Statistical inference and analysis of such large scale data sets is crucial in order to quantify statistical correctness of parameter estimates $\hat{\theta}_n$ (e.g., via Confidence Intervals) and testing hypotheses.

The Problem: Performing statistical inference on massive data sets is not computationally feasible using the conventional statistical inference methodology such as bootstrap.

Data Analysis problem at hand

- Confidence intervals may be more useful information than point estimate for Big Data.
- We develop scalable, robust and computationally efficient bootstrap technique for computing confidence intervals for big data.

CONTRIBUTION

A new bootstrap method is proposed. It facilitates bootstrap analysis of very large scale data. It is suited for estimators that can be expressed as a solution to fixed-point equations (e.g. M-estimator, MM-estimator, S-estimator, FastICA estimator).

We proof statistical convergence and quantitative robustness

The proposed method is:

1. Scalable to very large volume and high-dimensional data sets (Big Data).
2. Compatible with distributed data storage systems and distributed and parallel processing architectures.
3. Fast to compute as the fixed-point estimating equations do not need to be (re)solved for each bootstrap sample.
4. Statistically highly robust against outliers.

OUTLINE

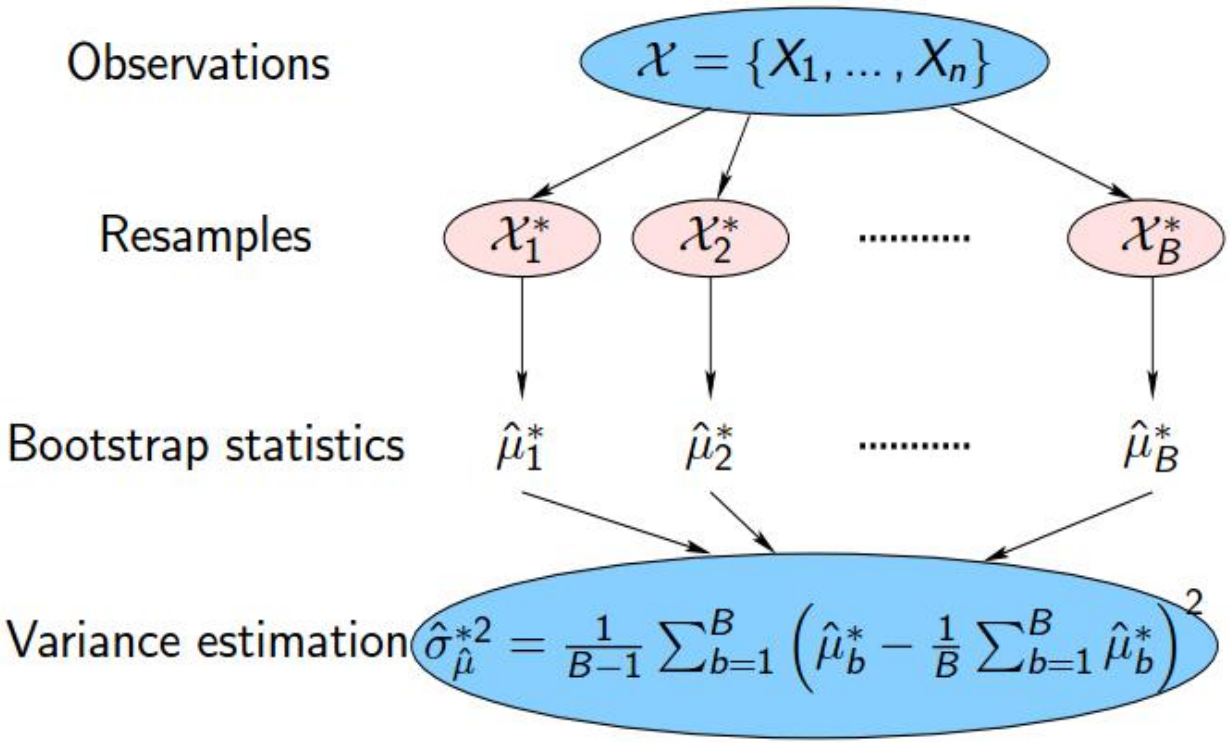
Bootstrap principles

- bootstrapping can refer to any test or metric that relies on random sampling from observed data with replacement.
  - resampling the observed data and performing inference on resampled replicas of the original dataset.
- Bootstrapping allows assigning quantitative measures of accuracy such as bias, variance, confidence intervals, prediction error to sample estimates
- properties of an estimator characterized by measuring those properties when sampling from an approximating distribution
- Typical choice for an approximating distribution is the empirical distribution function of the observed data.

Bootstrap principles

- If the data are i.i.d. approximation may be done by constructing a number of resamples with replacement of observed that are of equal size to the observed dataset.
- It may also be used for hypothesis testing. Statistical inference may be performed without assuming an explicit parametric model or when assumptions do not necessarily hold, or where parametric inference is impossible or very tedious.
- bootstrap works by treating inference of the true probability distribution F, given the original data, as being analogous to inference of the empirical distribution of F_e, given the resampled data.

Example: variance estimation (A. Zoubir)



When to use bootstrap?

- When the underlying theoretical distribution of an estimator or test of interest is complicated or unknown.
  - the bootstrap procedure does not assume distribution and provides an indirect method to assess the properties of the distribution underlying the observed data and the parameters of interest.
- When there are too few data for conventional statistical inference
- When calculating the sensitivity (power) of a binary hypothesis test and a small pilot dataset is available

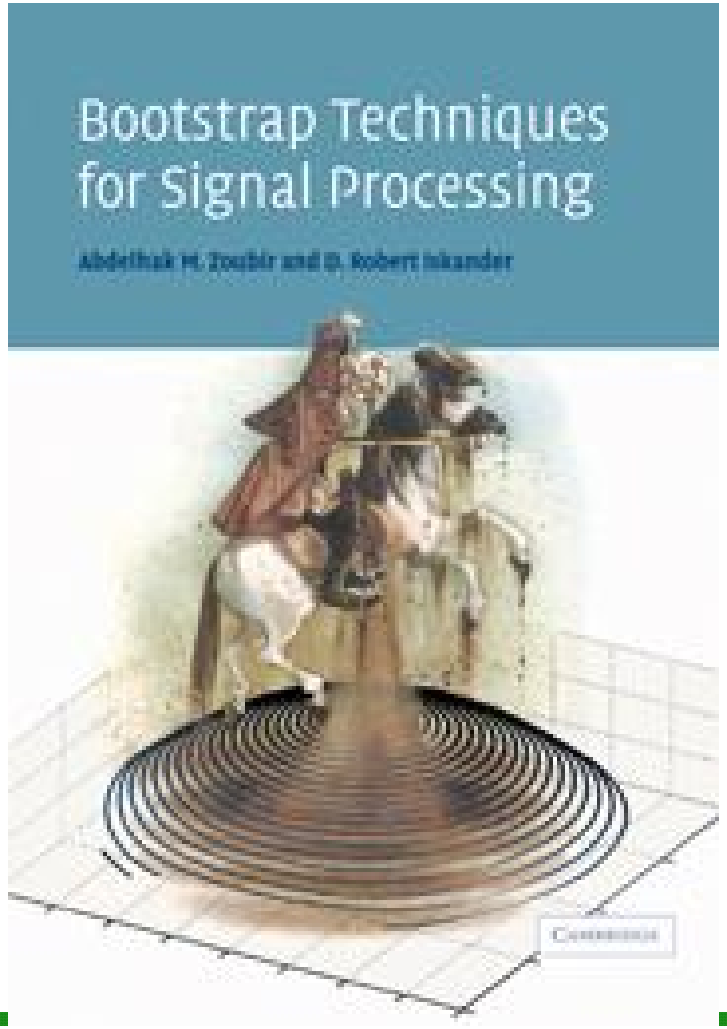
Bootstrap – pros and cons

- Advantages of bootstrap:
  - Simplicity, provides straightforward way to derive estimates of standard errors and confidence intervals even for complicated estimators of parameters of the distribution, such as percentile points and correlation coefficients.
- Bootstrap allows for controlling and checking the stability of the results.
- In many problems it is impossible to know the true confidence interval, bootstrap is asymptotically more accurate than the intervals obtained using sample variance and relying on assumption of normality

Bootstrap – pros and cons

- bootstrapping is often asymptotically consistent but it does not provide general finite-sample guarantees.
- Key assumptions such as independence of samples are implicitly made but not explicitly stated when undertaking the bootstrap analysis. In typical analytical derivations these would be more formally stated .

Excellent reference book



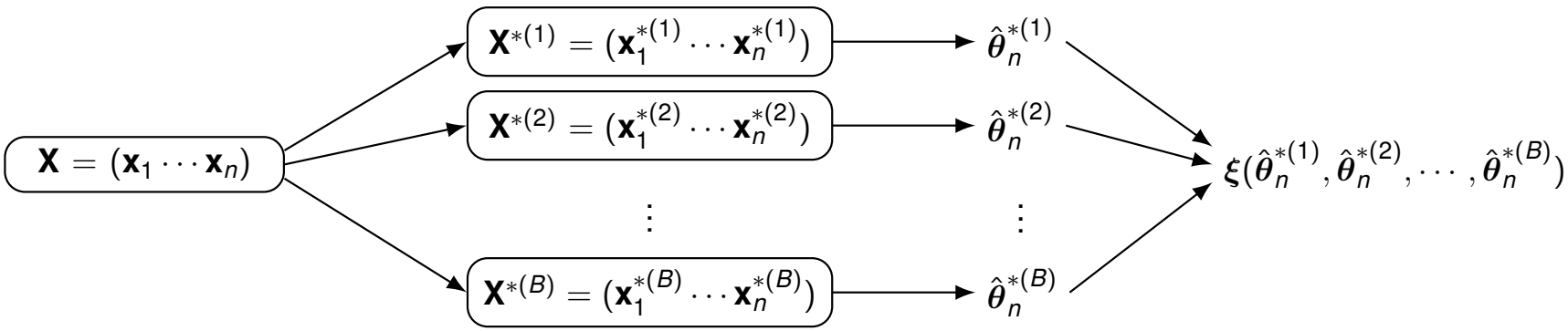
The Conventional Bootstrap

- The Bootstrap method (Efron) combines statistics and high-speed computational techniques, storage and resampling in order to find properties of estimators.
- Computational capabilities and resampling are used to compensate for the lack of knowledge on statistical properties or underlying distribution models.

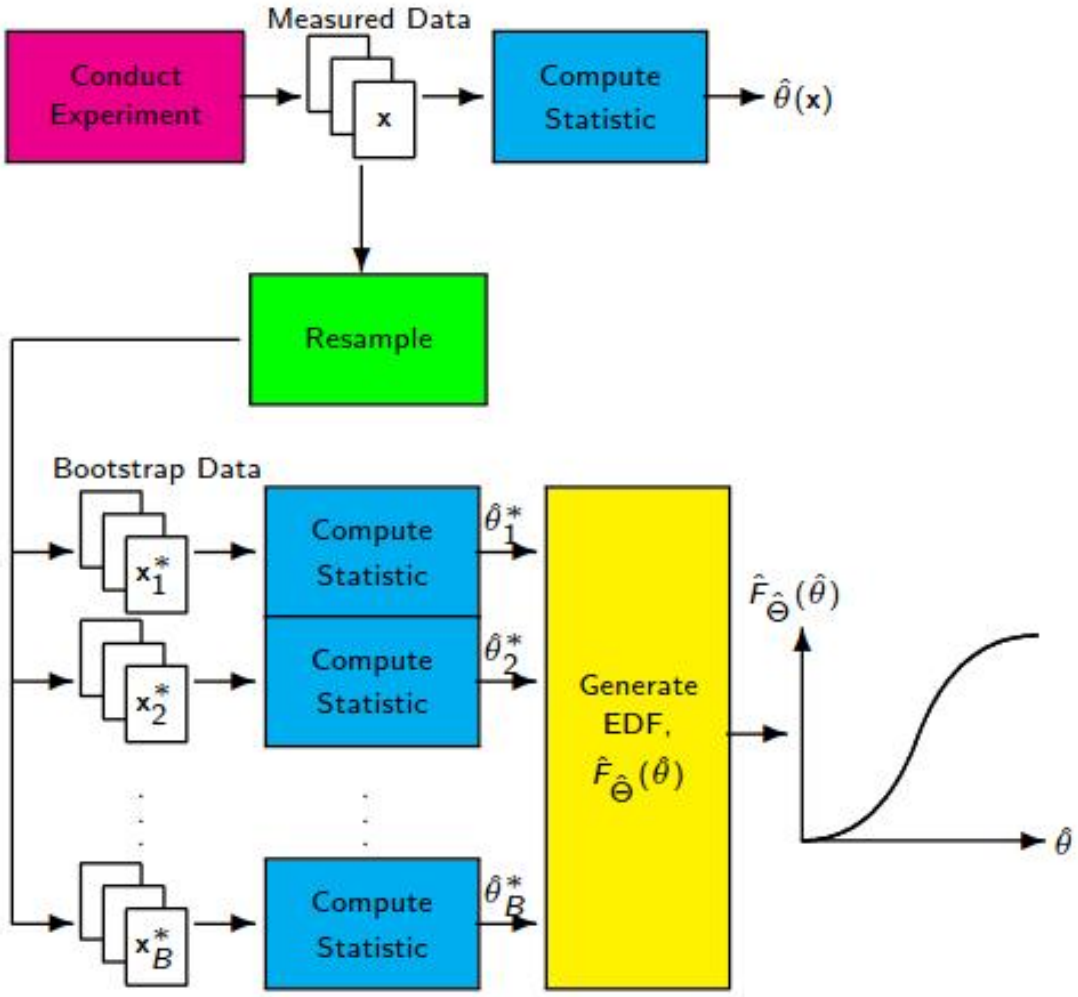
The Conventional Bootstrap

The bootstrap method [Efron, 1979] is a consistent and reliable method of constructing confidence intervals for statistical estimates (e.g., by bootstrap percentile method, BCA method, etc,).

1. Generate r bootstrap samples $\mathbf{X}^*$ of size n by resampling with replacement from the original data set $\mathbf{X}$.
2. Compute $\hat{\theta}_n^*$ on each bootstrap sample $\mathbf{X}^*$.
3. Use the population of bootstrap replications $\hat{\theta}_n^*$ to estimate the desired confidence intervals ξ.



Bootstrap


Measured Data
Conduct Experiment
x
Compute Statistic
$\hat{\theta}(\mathbf{x})$
Resample
Bootstrap Data
$\mathbf{x}_1^*$
$\mathbf{x}_2^*$
$\mathbf{x}_B^*$
Compute Statistic
Compute Statistic
Compute Statistic
$\hat{\theta}_1^*$
$\hat{\theta}_2^*$
$\hat{\theta}_B^*$
Generate EDF, $\hat{F}_{\hat{\Theta}}(\hat{\theta})$
$\hat{F}_{\hat{\Theta}}(\hat{\theta})$
$\hat{\theta}$

Remarks on computation

- The simplest method of finding confidence intervals for an unknown parameter is to take $\alpha/2$ and $1-\alpha/2$ quantiles of the bootstrap distribution of the estimator $\hat{\theta}_n$ as endpoints of the $100(1-\alpha)\%$ confidence interval
- Simple example: If we have 100 bootstrap estimates of θ, we will rank order them and the 90% confidence interval is found by choosing the 5th smallest and 95th value as the end points of the interval.
- In other words, trim $\alpha/2\%$ of the smallest and largest values off to find the confidence interval.

Remarks on computation

- Expected number of distinct datapoints in each resample is about $0.632n$
- Computational complexity typically scales with the number of distinct data points
- For a dataset of 1 TB, resample size would be 632 GB
- Parameter estimation and quality assesment is done for each resample
- Methods for reducing the number of resamples (Efron) and subsampling methods (m out of n bootstrap) have been proposed (Bickel)

The Conventional Bootstrap

- **Advantages:** Accurate for a wide range of estimators $\hat{\theta}_n$
- Automatic, because does not require any manipulation of the estimation equations
- **Disadvantages:** Computationally very costly since the estimator $\hat{\theta}_n$ is recomputed for each bootstrap sample
- Not robust in the face of outliers (highly deviating data)
- Not scalable, not suitable for distributed storage and (parallel) computing architectures.

The Conventional Bootstrap

Two main problems of using the bootstrap in analyzing large scale multivariate data sets:

1. The size of each bootstrap sample is the same as the original big data set (i.e., due to resampling with replacement, about 63% of data points appear at least once in each bootstrap sample).
2. Computation of the value of the estimator for each massive bootstrapped data set is not feasible using single storage and processing units.

The m out of n Bootstrap

The m out of n bootstrap [Bickel, et al., 1997] aims at reducing the computational cost by utilizing bootstrap samples of significantly smaller size than the original data set ($m < n$). The method is not suited for analysis of large multivariate data sets since:

- The output is sensitive to the size of the subsamples m.
- The knowledge of convergence rate of the estimator is needed in order to re-scale the output to the right size.
- The computational gains of the method (i.e., achieved by using smaller bootstrap samples) are reduced by the tedious analytical derivations needed for each inference problem at hand.

The Bag of Little Bootstraps (BLB)

The bag of little bootstraps (BLB) is a newly proposed bootstrap scheme [Kleiner, Jordan, et al., 2014] aiming to make the naive bootstrap method suitable for analysis of Big Data sets. In this method:

- Disjoint subsamples of significantly smaller size $b = \{\lfloor n^{\gamma} \rfloor \,|\gamma \in [0.6, 0.9]\}$ are drawn from the original Big Data set. Subsamples may be kept in/sent to in distributed storage and processing units for parallel computations.
- In each unit, bootstrap samples are constructed by assigning a random weight vector $\mathbf{n}^* = (n_1^*, \ldots, n_b^*)$ from $Multinomial(n, (1/b)\mathbf{1}_b)$ to the distinct data points of the subsample, where $\sum_{i=1}^{b} n_i^* = n$.

The BLB procedure

1. Draw s subsamples $\check{\mathbf{X}} = (\check{\mathbf{x}}_1 \ \cdots \ \check{\mathbf{x}}_b)$ of smaller size $b = \{\lfloor n^\gamma \rfloor \,|\gamma \in [0.6, 0.9]\}$ by randomly sampling *without* replacement from $\mathbf{X}$.

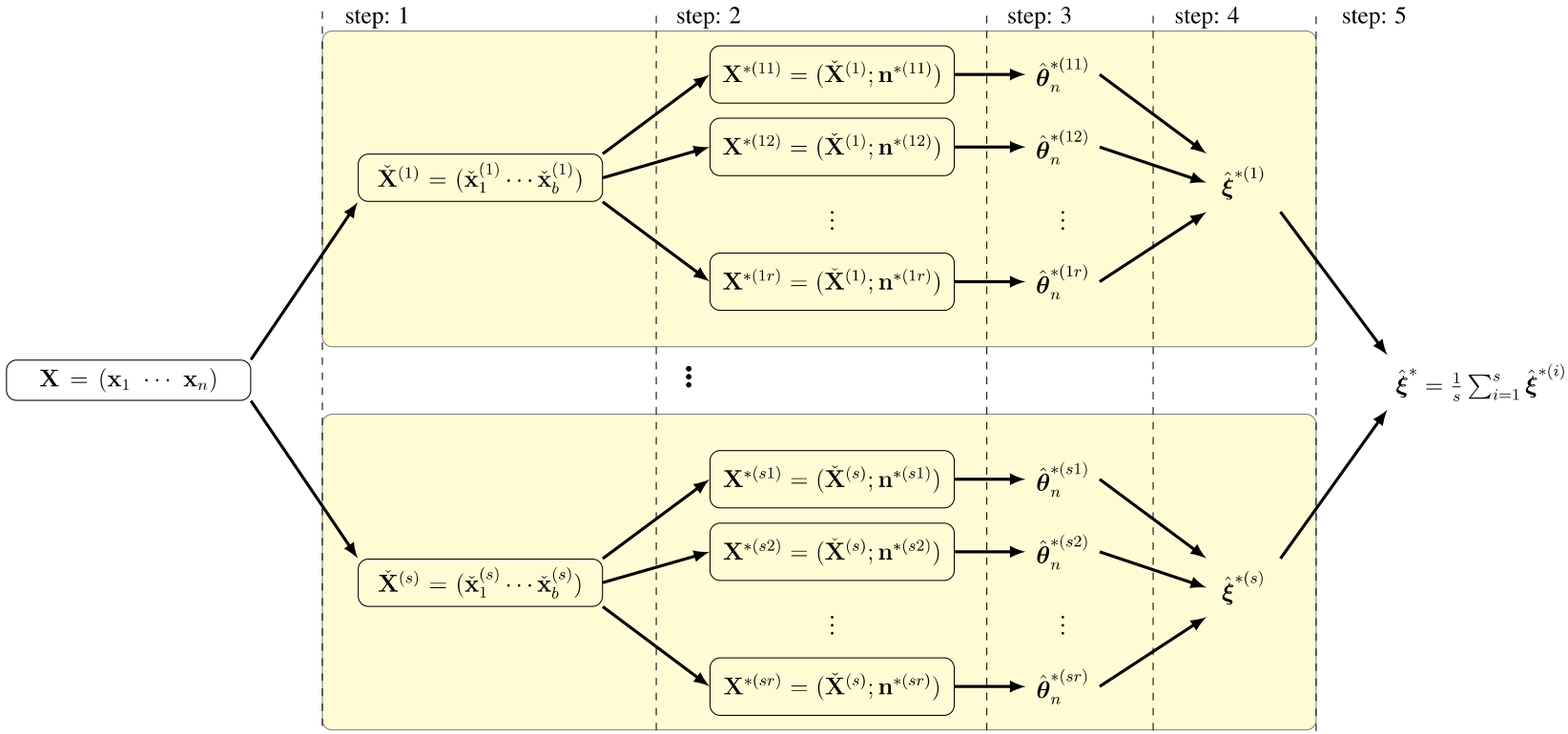
for each subsample $\check{\mathbf{X}}$

2. Generate r bootstrap samples $\mathbf{X}^* = (\check{\mathbf{X}}; \mathbf{n}^*)$ by assigning a random weight vector $\mathbf{n}^* = (n_1^*, \ldots, n_b^*)$ from $Multinomial(n, (1/b)\mathbf{1}_b)$ to data points of $\check{\mathbf{X}}$.
3. Compute the estimator $\hat{\theta}_n^*$ based on each $\mathbf{X}^*$.
4. Use the population of r bootstrap replications $\hat{\theta}_n^*$ to estimate the bootstrap confidence interval $\hat{\xi}^*$ (e.g., by bootstrap percentile method).

end

5. Average the computed values of $\hat{\xi}^*$ over the subsamples, i.e., $\hat{\xi}^* = \frac{1}{s}\sum_{k=1}^{s} \hat{\xi}^{*(k)}$.

The Bag of Little Bootstraps (BLB)



- $\check{\mathbf{X}}^{(k)}$, $k = 1, \ldots, s$ denotes the disjoint subsamples
- $\mathbf{X}^{*(kj)}$, $j = 1, \ldots, r$ corresponds to the jth BLB sample generated based on the subsample k.

The Bag of Little Bootstraps (Remarks)

- The distinct data of the subsamples in Step 2 allows the original Big data set to be stored in distributed storage systems.
- In Step 2, subsamples $\check{\mathbf{X}} = (\check{\mathbf{x}}_1 \ \cdots \ \check{\mathbf{x}}_b)$ can be processed in parallel using different computing nodes.
- $\mathbf{X}^* = (\check{\mathbf{X}}; \mathbf{n}^*)$ resembles a conventional bootstrap sample of size n with at most $b = \{\lfloor n^\gamma \rfloor \,|\gamma \in [0.6, 0.9]\}$ distinct data points. Element n_i^* of $\mathbf{n}^* = (n_1^*, \ldots, n_b^*)$ denotes the multiplicity of original subsample data point $\check{\mathbf{x}}_i$ at the bootstrap sample $\mathbf{X}^*$.
- BLB is computationally less complex than the conventional bootstrap. E.g., in the BLB scheme, $\mathbb{E}_{F_n^*}[\mathbf{X}^*]$ is computed by b *summations* $(+)$ and b *multiplications* $(\times)$ as $\mathbb{E}_{F_n^*}[\mathbf{X}^*] = \frac{1}{n}\sum_{i=1}^{b} n_i^* \check{\mathbf{x}}_i$, whereas in conventional method n *summations* $(+)$ are needed as $\mathbb{E}_{F_n^*}[\mathbf{X}^*] = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^*$.

The Bag of Little Bootstraps (BLB)

Advantages

- In comparison with the conventional bootstrap, less computational resources are needed.
- BLB is scalable and well suited for distributed computing architectures and storage systems.

Disadvantages

- The estimating equations need to be (re)solved for all bootstrap samples of all bags (overall $s \times r$ times). This is prohibitively expensive especially when a full optimization problem need to be numerically solved (e.g. matrix inversion or fixed-point iterative algorithm).
- The method is not statistically robust, in the sense that outlier contamination of only one subsample ruins the end result of the whole scheme.

The Fast and Robust Bootstrap (FRB)

FRB method [Salibian-Barrera, et al., 2008] is applicable for estimators $\hat{\boldsymbol{\theta}}_n \in \mathbb{R}^d$ that can be expressed as a solution to a system of smooth Fixed Point equations

$$\hat{\boldsymbol{\theta}}_n = \mathcal{Q}(\hat{\boldsymbol{\theta}}_n; \mathbf{X}), \tag{1}$$

where $\mathcal{Q} : \mathbb{R}^d \to \mathbb{R}^d$ and $\mathcal{Q}()$ is continuous and differentiable
The bootstrap replicated estimator $\hat{\boldsymbol{\theta}}^*_n$ then solves

$$\hat{\boldsymbol{\theta}}^*_n = \mathcal{Q}(\hat{\boldsymbol{\theta}}^*_n; \mathbf{X}^*), \tag{2}$$

Fixed point equations usually converge fast, in few iterations.

The Fast and Robust Bootstrap (FRB)

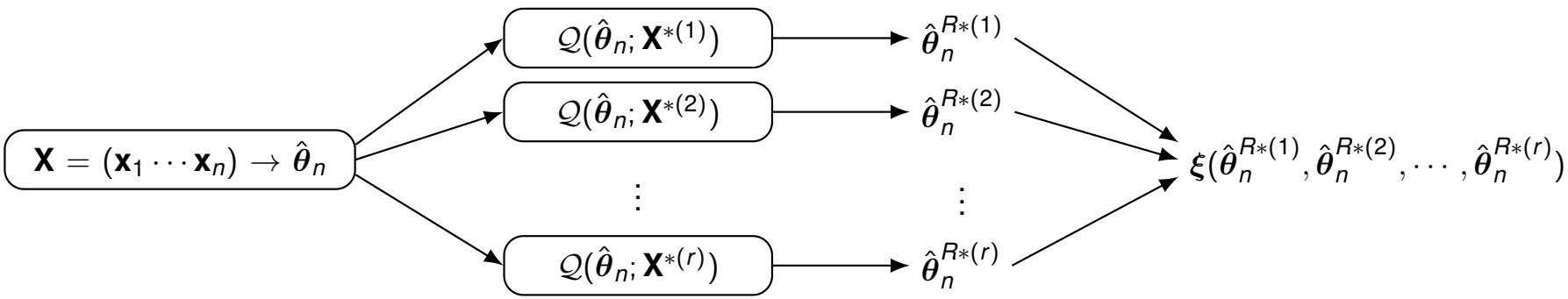
Instead of computing $\hat{\theta}_n^*$, we compute an approximation:

$$\hat{\theta}_n^{1*} = \mathcal{Q}(\hat{\theta}_n; \mathbf{X}^*), \tag{3}$$

Since the distribution of $\hat{\theta}_n^{1*}$ typically underestimates the sampling variability of $\hat{\theta}_n$, a linear correction based on Taylor approximation of function $\mathcal{Q}$ is applied as follows:

$$\hat{\theta}_n^{R*} = \hat{\theta}_n + \left[\mathbf{I} - \nabla\mathcal{Q}(\hat{\theta}_n; \mathbf{X})\right]^{-1}\left(\hat{\theta}_n^{1*} - \hat{\theta}_n\right), \tag{4}$$

where $\nabla\mathcal{Q}(\cdot) \in \mathbb{R}^{d\times d}$ is the matrix of partial derivatives w.r.t. $\hat{\theta}_n$.



The Fast and Robust Bootstrap (FRB)

Advantages

- Fast to compute, as the initial estimator $\hat{\theta}_n$ is computed only once (e.g, for the full data set **X**). One step improvement $\hat{\theta}_n^{1*}$ requires only one iteration of FP equation.
- Robust against outliers. For instance in case of the MM-regression estimator, it has been shown that equation (4) remains bounded if $\hat{\theta}_n$ is a reliable estimate of θ and there are only p (the dimension of the regression model) non-outlier data points in the bootstrap sample $\mathbf{X}^*$.

Disadvantages

- Not scalable and difficult to parallelize across distributed computing systems.$\hat{\theta}_n$ computed from large scale data **X**

Simple example formulation for M-estimator of linear regression

Let $\mathbf{X} = \{(y_1, \mathbf{z}_1^\top)^\top, \ldots, (y_n, \mathbf{z}_n^\top)^\top\}$, $\mathbf{z}_i \in \mathbb{R}^p$, be a sample of independent random vectors that follow the linear model:

$$y_i = \mathbf{z}_i^\top \boldsymbol{\theta} + \sigma_0 e_i \quad \text{for} \quad i = 1, \ldots, n, \tag{5}$$

where:

- $\boldsymbol{\theta} \in \mathbb{R}^p$: The unknown parameter vector.
- e_i : Noise terms are i.i.d. random variables from a symmetric distribution with unit scale.

Simple example formulation FRB for M-estimator of linear regression

The M-estimators of regression are defined as:

$$\frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{y_i-\mathbf{z}_i^\top\hat{\boldsymbol{\theta}}_n}{\hat{\sigma}_n}\right)\mathbf{z}_i=\mathbf{0}, \tag{6}$$

where the scale parameter $\hat{\sigma}_n$ needs to be estimated from the data. M-estimator is obtained by generalizing the maximum likelihood estimator (MLE) such that the ψ function need not be related to any particular error density, but it can be any continuous, bounded and odd function.

Example formulation for M-estimator of linear regression

Equation (6) can be expressed in form of a fixed-point estimating equation:

$$\hat{\boldsymbol{\theta}}_n = \mathcal{Q}(\hat{\boldsymbol{\theta}}_n; \mathbf{Z}) = (\mathbf{Z}^\top \mathbf{W} \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{W} \mathbf{y}, \tag{7}$$

where

$$\mathbf{Z}^\top = (\mathbf{z}_1, \cdots, \mathbf{z}_n) \in \mathbb{R}^{p \times n}, \qquad \mathbf{y} = (y_1, \cdots, y_n)^\top \in \mathbb{R}^{n \times 1},$$
$$\mathbf{W} = \operatorname{diag}\{\omega_1, \cdots, \omega_n\},$$
$$\omega_i = \psi(r_i / \hat{\sigma}_n)/r_i \qquad \text{and} \qquad r_i = y_i - \mathbf{z}_i^\top \hat{\boldsymbol{\theta}}_n,$$
$$\text{for} \quad i = 1, \ldots, n.$$

Example formulation for M-estimator of linear regression

For a given bootstrap sample $\mathbf{X}^* = \{(y_1^*, \mathbf{z}_1^{*\top})^\top, \ldots, (y_n^*, \mathbf{z}_n^{*\top})^\top\}$, the one-step iteration of the fixed-point equation becomes:

$$\hat{\theta}_n^{1*} = \mathcal{Q}(\hat{\theta}_n; \mathbf{X}^*) = (\mathbf{Z}^{*\top}\mathbf{W}^*\mathbf{Z}^*)^{-1}\mathbf{Z}^{*\top}\mathbf{W}^*\mathbf{y}^*, \tag{8}$$

where

$$\mathbf{Z}^{*\top} = (\mathbf{z}_1^*, \cdots, \mathbf{z}_n^*) \in \mathbb{R}^{p\times n}, \qquad \mathbf{y}^* = (y_1^*, \cdots, y_n^*)^\top \in \mathbb{R}^{n\times 1},$$
$$\mathbf{W}^* = \text{diag}\{\omega_1^*, \cdots, \omega_n^*\},$$
$$\omega_i^* = \psi(r_i^*/\hat{\sigma}_n^*)/r_i^* \qquad \text{and} \qquad r_i^* = y_i^* - \mathbf{z}_i^{*\top}\hat{\theta}_n,$$
$$\text{for} \quad i = 1, \ldots, n.$$

Example formulation for M-estimator of linear regression

The FRB replication of $\hat{\theta}_n$ is obtained by applying the correction term on equation (8).

$$\hat{\theta}_n^{R*} = \hat{\theta}_n + \left[\mathbf{I} - \nabla\mathcal{Q}(\hat{\theta}_n; \mathbf{X})\right]^{-1}(\hat{\theta}_n^{1*} - \hat{\theta}_n), \tag{9}$$

where $\nabla\mathcal{Q}(\cdot) \in \mathbb{R}^{d\times d}$ is the matrix of partial derivatives of the fixed-point equation (7) w.r.t. $\hat{\theta}_n$,

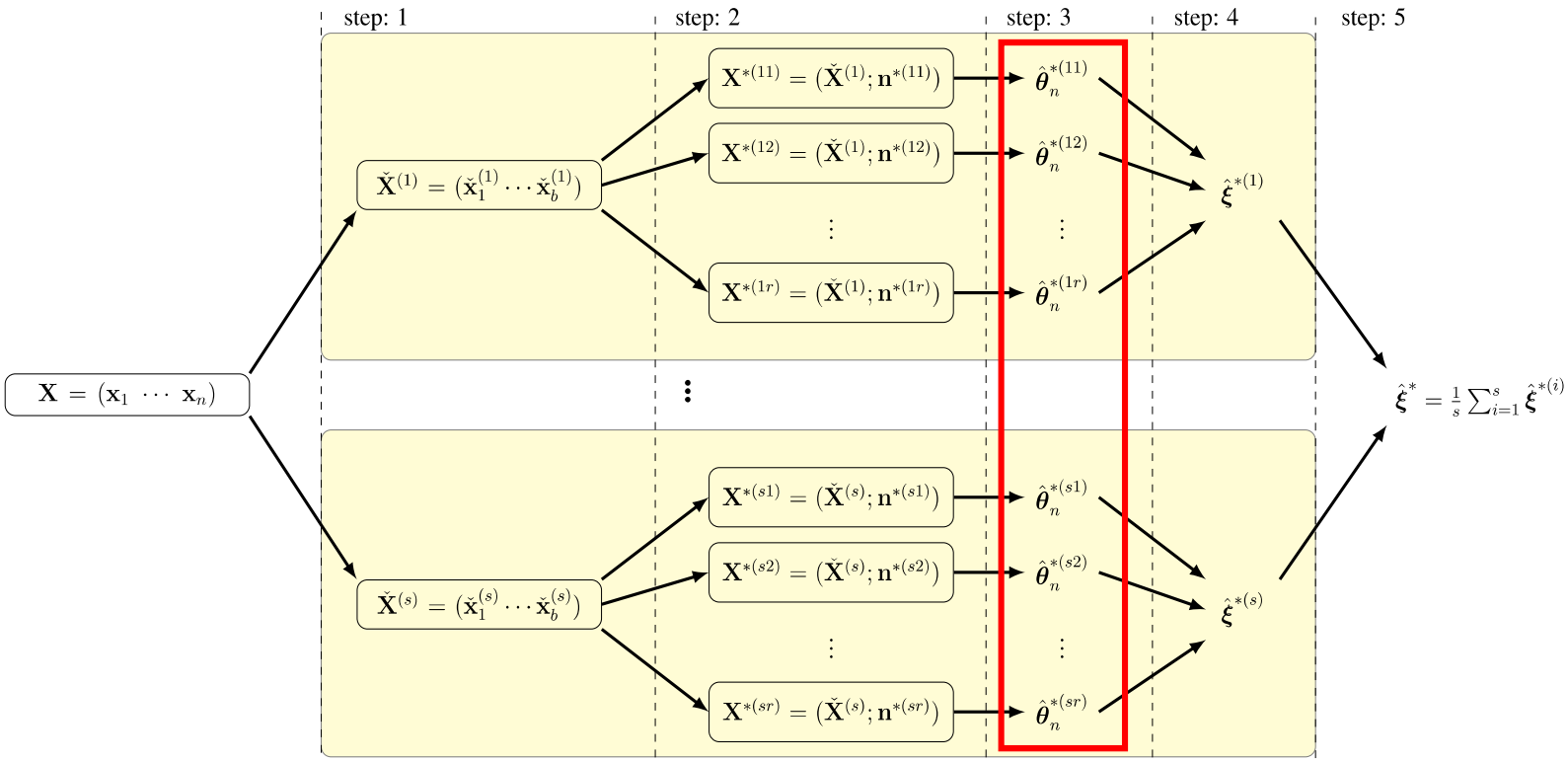
$$\nabla\mathcal{Q}(\hat{\theta}_n; \mathbf{X}) = \frac{\partial(\mathbf{Z}^\top\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{W}\mathbf{y}}{\partial\hat{\theta}_n}.$$

FAST AND ROBUST BOOTSTRAP FOR BIG DATA (BLFRB)

The main contribution is to introduce a new bootstrap scheme that is suitable for analyzing large multivariate data sets.
The BLFRB method [S. Basiri, E. Ollila, V. Koivunen, 2015] combines the desirable properties of the BLB and FRB methods as it is:

1. Scalable to large volume data sets and compatible with distributed data storage and processing architectures.
2. Less complex and fast to compute as the estimating equations are computed only once for each bag.
3. Statistically robust and works reliably in the face of outliers. Bootstrap analysis of fixed-point robust estimators (e.g. S-estimator, MM-estimator, etc.) is facilitated, thanks to the low complexity of the scheme.

FAST AND ROBUST BOOTSTRAP FOR BIG DATA



Recall that the main computational burden of the BLB scheme is in step 3 of the algorithm where, the estimating equations need to be (re)solved for each bootstrap sample X^*.

FAST AND ROBUST BOOTSTRAP FOR BIG DATA

- Such computational complexity can be drastically reduced by computing the FRB replications instead. This can be done locally within each bag:
  - Let $\hat{\boldsymbol{\theta}}_{n,b}$ be a solution to $\hat{\boldsymbol{\theta}}_n = \mathcal{Q}(\hat{\boldsymbol{\theta}}_n; \mathbf{X})$, for subsample $\check{\mathbf{X}} \in \mathbb{R}^{d \times b}$:
  $$\hat{\boldsymbol{\theta}}_{n,b} = \mathcal{Q}(\hat{\boldsymbol{\theta}}_{n,b}; \check{\mathbf{X}}). \tag{10}$$
  - Let $\mathbf{X}^* \in \mathbb{R}^{d \times n}$ be a bootstrap sample of size n randomly resampled with replacement from distinct data subset $\check{\mathbf{X}}$ of size b;
  - The FRB replication of $\hat{\boldsymbol{\theta}}_{n,b}$ can be obtained by
  $$\hat{\boldsymbol{\theta}}^{R*}_{n,b} = \hat{\boldsymbol{\theta}}_{n,b} + \left[\mathbf{I} - \nabla\mathcal{Q}(\hat{\boldsymbol{\theta}}_{n,b}; \check{\mathbf{X}})\right]^{-1} \left(\hat{\boldsymbol{\theta}}^{1*}_{n,b} - \hat{\boldsymbol{\theta}}_{n,b}\right), \tag{11}$$
  where $\hat{\boldsymbol{\theta}}^{1*}_{n,b} = \mathcal{Q}(\hat{\boldsymbol{\theta}}_{n,b}; \mathbf{X}^*)$ is the one-step estimator and $\nabla\mathcal{Q}(\cdot) \in \mathbb{R}^{d \times d}$ is the matrix of partial derivatives w.r.t. $\hat{\boldsymbol{\theta}}_{n,b}$.
- Note: The initial estimate $\hat{\boldsymbol{\theta}}_{n,b}$ and correction $\left[\mathbf{I} - \nabla\mathcal{Q}(\hat{\boldsymbol{\theta}}_{n,b}; \check{\mathbf{X}})\right]^{-1}$ are computed only once for each distinct data subset.

THE PROPOSED BLFRB PROCEDURE

1: Draw s disjoint subsamples $\check{\mathbf{X}} = (\check{\mathbf{x}}_1 \ \cdots \ \check{\mathbf{x}}_b)$ of smaller size $b = \{\lfloor n^\gamma \rfloor \,|\gamma \in [0.6, 0.9]\}$.

for each subsample $\check{\mathbf{X}}$

2: Generate r bootstrap samples $\mathbf{X}^* = (\check{\mathbf{X}}; \mathbf{n}^*)$ according to the BLB procedure.

3:
- a: Find the estimate $\hat{\theta}_{n,b}$ based on $\check{\mathbf{X}}$.
- b: For each bootstrap sample $\mathbf{X}^*$ compute the FRB replication $\hat{\theta}^{R*}_{n,b}$ using $\hat{\theta}_{n,b}$.

4: Compute the bootstrap confidence intervals $\hat{\xi}^*$ based on the population of r FRB replicated values $\hat{\theta}^{R*}_{n,b}$.

end

5: Average the computed values of $\hat{\xi}^*$ over the subsamples, i.e., $\hat{\xi}^* = \frac{1}{s}\sum_{k=1}^{s} \hat{\xi}^{*(k)}$.

Example formulation of BLFRB for MM-estimator of linear regression

Let $\mathbf{X} = \{(y_1, \mathbf{z}_1^\top)^\top, \ldots, (y_n, \mathbf{z}_n^\top)^\top\}$, $\mathbf{z}_i \in \mathbb{R}^p$, be a sample of independent random vectors that follow the linear model:

$$y_i = \mathbf{z}_i^\top \boldsymbol{\theta} + \sigma_0 e_i \quad \text{for} \quad i = 1, \ldots, n, \tag{12}$$

where:

- $\boldsymbol{\theta} \in \mathbb{R}^p$: The unknown parameter vector.
- e_i : Noise terms are i.i.d. random variables from a symmetric distribution with unit scale.

Example formulation of BLFRB for MM-estimator of linear regression

Highly robust MM-estimators [V. J. Yohai, 1987] proceed in 3 stages: (1) initial highly robust estimate that is not necessarily efficient is found; (2) M-estimate of error scale is computed based on residuals; (3) M-estimate of the parameter vector is computed.

Two loss functions $\rho_0 : \mathbb{R} \to \mathbb{R}^+$ and $\rho_1 : \mathbb{R} \to \mathbb{R}^+$ are used which determine the breakdown point and efficiency of the estimator. The $\rho_0(\cdot)$ and $\rho_1(\cdot)$ functions are

- Symmetric,
- Twice continuously differentiable with $\rho(0) = 0$,
- Strictly increasing on $[0, c]$ and constant on $[c, \infty)$ for some constant c.

Example formulation of BLFRB for MM-estimator of linear regression

The MM-estimate of $\hat{\boldsymbol{\theta}}_n$ satisfies

$$\frac{1}{n}\sum_{i=1}^{n}\rho_1'\left(\frac{y_i-\mathbf{z}_i^\top\hat{\boldsymbol{\theta}}_n}{\hat{\sigma}_n}\right)\mathbf{z}_i=\mathbf{0} \tag{13}$$

where $\hat{\sigma}_n$ is a S-estimate of scale defined as follows.

- Consider M-estimate of scale $\hat{s}_n(\boldsymbol{\theta})$ obtained as a solution to:

$$\frac{1}{n}\sum_{i=1}^{n}\rho_0\left(\frac{y_i-\mathbf{z}_i^\top\boldsymbol{\theta}}{\hat{s}_n(\boldsymbol{\theta})}\right)=m, \tag{14}$$

where $m=\rho_0(\infty)/2$ is a constant. Let $\tilde{\boldsymbol{\theta}}_n$ be the argument that minimizes $\hat{s}_n(\boldsymbol{\theta})$,

$$\tilde{\boldsymbol{\theta}}_n=\arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p}\hat{s}_n(\boldsymbol{\theta}),$$

then $\hat{\sigma}_n=\hat{s}_n(\tilde{\boldsymbol{\theta}}_n)$.

Example formulation of BLFRB for MM-estimator of linear regression

Simple computations yield the following FP representation of (13) and (14):

$$\hat{\boldsymbol{\theta}}_n = \left(\sum_{i=1}^{n} \omega_i \mathbf{z}_i \mathbf{z}_i^\top \right)^{-1} \sum_{i=1}^{n} \omega_i \mathbf{z}_i y_i, \tag{15}$$

$$\hat{\sigma}_n = \sum_{i=1}^{n} \upsilon_i (y_i - \mathbf{z}_i^\top \tilde{\boldsymbol{\theta}}_n), \tag{16}$$

where

$$r_i = y_i - \mathbf{z}_i^\top \hat{\boldsymbol{\theta}}_n, \qquad \dot{r}_i = y_i - \mathbf{z}_i^\top \tilde{\boldsymbol{\theta}}_n,$$

$$\omega_i = \rho_1'(r_i/\hat{\sigma}_n)/r_i \quad \text{and} \quad \upsilon_i = \frac{\hat{\sigma}_n}{nm} \rho_0(\dot{r}_i/\hat{\sigma}_n)/\dot{r}_i.$$

Example formulation of BLFRB for MM-estimator of linear regression

Let $\mathbf{X}^* = (\check{\mathbf{X}}; \mathbf{n}^*)$ denote a BLFRB bootstrap sample based on subsample $\check{\mathbf{X}} = \{(\check{y}_1, \check{\mathbf{z}}_1^\top)^\top, \ldots, (\check{y}_b, \check{\mathbf{z}}_b^\top)^\top\}$, $\check{\mathbf{z}}_i \in \mathbb{R}^p$ and a weight vector $\mathbf{n}^* = (n_1^* \cdots n_b^*) \in \mathbb{R}^b$,

$$\hat{\boldsymbol{\theta}}_{n,b}^{1*} = \left(\sum_{i=1}^{b} n_i^* \check{\omega}_i \check{\mathbf{z}}_i \check{\mathbf{z}}_i^\top \right)^{-1} \sum_{i=1}^{b} n_i^* \check{\omega}_i \check{\mathbf{z}}_i \check{y}_i, \tag{17}$$

$$\hat{\sigma}_{n,b}^{1*} = \sum_{i=1}^{b} n_i^* \check{\upsilon}_i (\check{y}_i - \check{\mathbf{z}}_i^\top \tilde{\boldsymbol{\theta}}_{n,b}), \tag{18}$$

where

$$\check{r}_i = \check{y}_i - \check{\mathbf{z}}_i^\top \hat{\boldsymbol{\theta}}_{n,b}, \qquad \tilde{r}_i = \check{y}_i - \check{\mathbf{z}}_i^\top \tilde{\boldsymbol{\theta}}_{n,b},$$

$$\check{\omega}_i = \rho_1'(\check{r}_i / \hat{\sigma}_{n,b}) / \check{r}_i \quad \text{and} \quad \check{\upsilon}_i = \frac{\hat{\sigma}_{n,b}}{nm} \rho_0(\tilde{r}_i / \hat{\sigma}_{n,b}) / \tilde{r}_i.$$

Example formulation of BLFRB for MM-estimator of linear regression

The BLFRB replications of $\hat{\theta}_{n,b}$, are obtained from the linearly corrected version of the one-step approximations in (17) and (18):

$$\hat{\theta}^{R*}_{n,b} = \hat{\theta}_{n,b} + \mathbf{M}_{n,b}(\hat{\theta}^{1*}_{n,b} - \hat{\theta}_{n,b}) + \mathbf{d}_{n,b}(\hat{\sigma}^{1*}_{n,b} - \hat{\sigma}_{n,b}), \qquad (19)$$

where

$$\mathbf{M}_{n,b} = \hat{\sigma}_{n,b}\left(\sum_{i=1}^{b} \rho_1''(\check{r}_i/\hat{\sigma}_{n,b})\check{\mathbf{z}}_i\check{\mathbf{z}}_i^\top\right)^{-1}\sum_{i=1}^{b}\check{\omega}_i\check{\mathbf{z}}_i\check{\mathbf{z}}_i^\top,$$

$$\mathbf{d}_{n,b} = k_{n,b}^{-1}\left(\sum_{i=1}^{b} \rho_1''(\check{r}_i/\hat{\sigma}_{n,b})\check{\mathbf{z}}_i\check{\mathbf{z}}_i^\top\right)^{-1}\sum_{i=1}^{b}\rho_1''(\check{r}_i/\hat{\sigma}_{n,b})\check{r}_i\check{\mathbf{z}}_i$$

and $$k_{n,b} = \frac{1}{nm}\sum_{i=1}^{b}\left(\rho_0'(\tilde{r}_i/\hat{\sigma}_{n,b})\tilde{r}_i/\hat{\sigma}_{n,b}\right).$$

STATISTICAL PROPERTIES

Statistical Convergence

Notation

- $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$: A set of observed data as the outcome of i.i.d. random variables $\boldsymbol{\mathcal{X}} = \{\boldsymbol{\mathcal{X}}_1, \ldots, \boldsymbol{\mathcal{X}}_n\}$ from an unknown distribution P.
- $\mathbb{P}_n = n^{-1} \sum_{i=1}^{n} \delta_{\mathbf{x}_i}$: The empirical distribution (measure) formed by $\mathbf{X}$.
- $\mathbb{P}_{n,b}^{(k)} = n^{-1} \sum_{i=1}^{b} \frac{n}{b} \delta_{\check{\mathbf{x}}_i^{(k)}}$: the empirical distribution formed by subsample $\check{\mathbf{X}}^{(k)}$.
- $\mathbb{P}_{n,b}^{*} = n^{-1} \sum_{i=1}^{n} \delta_{\mathbf{x}_i^{*}}$: The empirical distribution formed by bootstrap sample $\mathbf{X}^{*}$.
- $\phi(\cdot)$: Functional representations of the estimator e.g., $\boldsymbol{\theta} = \phi(P), \hat{\boldsymbol{\theta}}_{n,b}^{(k)} = \phi(\mathbb{P}_{n,b}^{(k)})$ and $\hat{\boldsymbol{\theta}}_{n,b}^{*} = \phi(\mathbb{P}_{n,b}^{*})$.
- $\overset{d}{=}$: Denotes that both sides have the same limiting distribution.

STATISTICAL PROPERTIES

Statistical Convergence

Theorem

Consider P, $\mathbb{P}_n$ and $\mathbb{P}_{n,b}^{(k)}$ as maps from a Donsker class $\mathcal{F}$ to $\mathbb{R}$ such that $\mathcal{F}_\delta = \{f - g : f, g \in \mathcal{F}, \{P(f - Pf)^2\}^{1/2} < \delta\}$ is measurable for every $\delta > 0$. Let ϕ to be Hadamard differentiable at P tangentially to some subspace and $\hat{\boldsymbol{\theta}}_n$ be a solution to a system of smooth FP equations. Then as $n, b \to \infty$

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{n,b}^{R*} - \hat{\boldsymbol{\theta}}_{n,b}^{(k)}) \stackrel{d}{=} \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}). \tag{20}$$

See the proof in [S. Basiri, E. Ollila, V. Koivunen, 2015].

STATISTICAL PROPERTIES

Statistical robustness

Notation:

Let $\mathbf{X} = \{(y_1, \mathbf{z}_1^\top)^\top, \ldots, (y_n, \mathbf{z}_n^\top)^\top\}$, $\mathbf{z}_i \in \mathbb{R}^p$, be a sample of *iid* random vectors that follow the linear model, $y_i = \mathbf{z}_i^\top \boldsymbol{\theta} + \sigma_0 e_i$ for $i = 1, \ldots, n$, and let

- $\hat{\boldsymbol{\theta}}_n$ be an estimator of the parameter vector $\boldsymbol{\theta}$ based on $\mathbf{X}$,
- $q_t, t \in (0, 1)$, denote the tth upper quantile of $[\hat{\boldsymbol{\theta}}_n]_l$, where $[\hat{\boldsymbol{\theta}}_n]_l$ is the lth element of $\hat{\boldsymbol{\theta}}_n$, $l = 1, \ldots, p$ e.g., $Pr\left([\hat{\boldsymbol{\theta}}_n]_l > q_t\right) = t$.
- $\hat{q}_t^*$ denote the BLB or BLFRB estimate of the q_t based on a random subsample $\check{\mathbf{X}}$ of size $b = \{\lfloor n^\gamma \rfloor \,|\gamma \in [0.6, 0.9]\}$ drawn from a big data set $\mathbf{X}$.

Definition: The upper breakdown point of $\hat{q}_t^*$ is defined as the minimum proportion of asymmetric outlier contamination in subsample $\check{\mathbf{X}}$ that can drive $\hat{q}_t^*$ over any finite bound.

STATISTICAL PROPERTIES

Statistical robustness

Theorem

In the original BLB setting with Least Square estimator, only one outlying data point in a subsample $\check{\mathbf{X}}$ *is sufficient to drive* $\hat{q}_t^*$, $t \in (0, 1)$ *over any finite bound and hence, ruining the end result of the whole scheme.*

See the proof in [S. Basiri, E. Ollila, V. Koivunen, 2015].

STATISTICAL PROPERTIES

Statistical robustness

Theorem

Let $\check{\mathbf{X}} = \{(\check{y}_1, \check{\mathbf{z}}_1^\top)^\top, \ldots, (\check{y}_b, \check{\mathbf{z}}_b^\top)^\top\}$*, be a subsample of size* $b = \{\lfloor n^\gamma \rfloor \,|\gamma \in [0.6, 0.9]\}$ *randomly drawn from* $\mathbf{X}$ *following the linear model* $y_i = \mathbf{z}_i^\top \boldsymbol{\theta} + \sigma_0 e_i$ *for* $i = 1, \ldots, n$*. Assume that the explaining variables* $\check{\mathbf{z}}_1^\top, \ldots, \check{\mathbf{z}}_b^\top \in \mathbb{R}^p$ *are in general position. Let* $\hat{\boldsymbol{\theta}}_{n,b}$ *be an MM-estimator of* $\boldsymbol{\theta}$ *based on* $\check{\mathbf{X}}$ *and let* δ_b *be the finite sample breakdown point of* $\hat{\boldsymbol{\theta}}_{n,b}$*. Then in the BLFRB bag formed by* $\check{\mathbf{X}}$*, all the estimated quantiles* $\hat{q}_t^*$*,* $t \in (0, 1)$ *have the same breakdown point equal to* δ_b*.*
See the proof in [S. Basiri, E. Ollila, V. Koivunen, 2015].

Note: The finite sample breakdown point of MM-estimator can be set close to 0.5. This provides the maximum possible statistical robustness for the quantile estimates.

STATISTICAL PROPERTIES

Statistical robustness

p	n	$\gamma = 0.6$	$\gamma = 0.7$	$\gamma = 0.8$
50	50000	0.425	0.475	0.491
	200000	0.467	0.490	0.497
	1000000	0.488	0.497	0.499
100	50000	0.349	0.449	0.483
	200000	0.434	0.481	0.494
	1000000	0.475	0.494	0.498
200	50000	0.197	0.398	0.465
	200000	0.368	0.461	0.488
	1000000	0.450	0.487	0.497

Table: Upper breakdown point of the BLFRB estimates of quantiles for MM-regression estimator with 50% breakdown point and 95% efficiency at the Gaussian model.

NUMERICAL EXAMPLES
The model

We generate $n = 50000$ samples $\mathbf{X} = \{(y_1, \mathbf{z}_1^\top)^\top, \ldots, (y_n, \mathbf{z}_n^\top)^\top\}$ from linear model with unknown parameters $\boldsymbol{\theta}$:

$$y_i = \mathbf{z}_i^\top \boldsymbol{\theta} + \sigma_0 e_i,$$

where:

- The explaining variables $\mathbf{z}_i$ have p-variate normal distribution $\mathcal{N}_p(\mathbf{0}, \mathbf{I}_p)$ with $p = 50$.
- The parameter vector $\boldsymbol{\theta} = \mathbf{1}_p$.
- The noise terms, e_i are i.i.d. from the standard normal distribution.
- The variance of the noise is $\sigma_0^2 = 0.1$.
- The MM-estimator in the BLFRB scheme is tuned to have efficiency $\mathcal{O} = 95\%$ and breakdown point $\delta = 50\%$.
- The original BLB scheme uses the Least Square-estimator for computation of the bootstrap estimates of $\boldsymbol{\theta}$.

NUMERICAL EXAMPLES

Statistical Convergence of BLFRB

Consider the result of the BLFRB convergence theorem:

$$\sqrt{n}(\hat{\theta}_{n,b}^{R*} - \hat{\theta}_{n,b}^{(k)}) \overset{d}{=} \sqrt{n}(\hat{\theta}_n - \theta).$$

- Given the above settings, the right hand side follows $\mathcal{N}_p(\mathbf{0}, \sigma_0^2/\mathcal{O}\mathbf{I}_p)$ in distribution [V. J. Yohai, 1987, theorem 4.1].
- We form the distribution of the left hand side, by drawing a random subsample $\check{\mathbf{X}}$ of size $b = \lfloor 50000^{0.7} \rfloor = 1946$ and performing steps 2 and 3 of the BLFRB procedure for $\check{\mathbf{X}}$ using $r = 1000$ bootstrap samples.

NUMERICAL EXAMPLES

Statistical Convergence of BLFRB

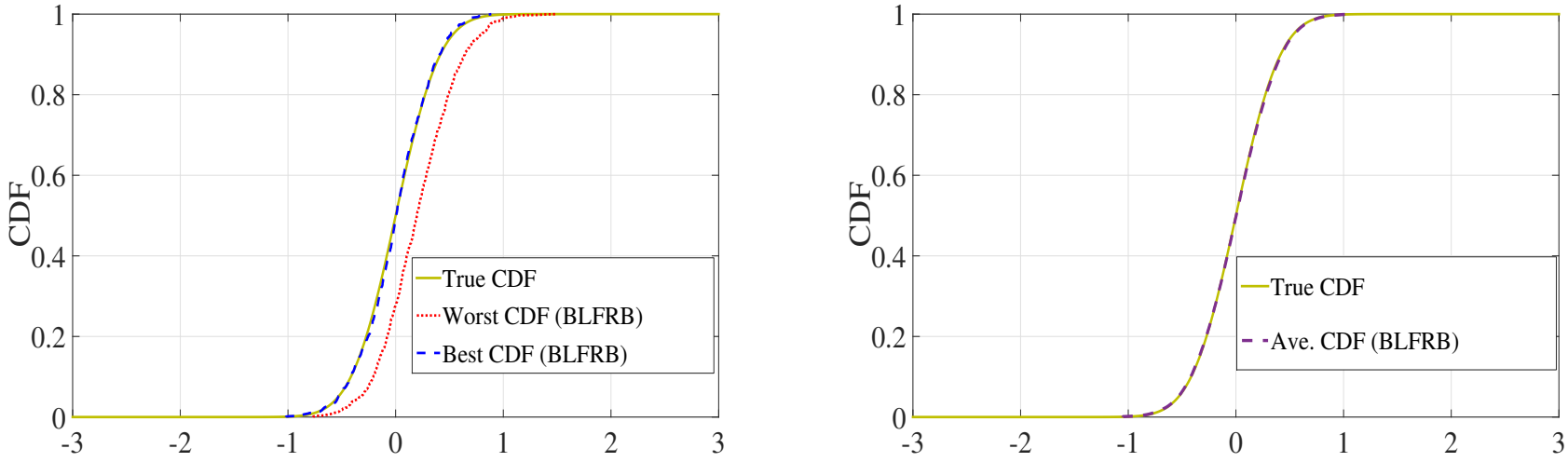


Figure: **(a)** The true distribution of the right hand side of (20) along with the obtained empirical distributions of the left hand side for two elements of $\hat{\theta}_{n,b}^{R*}$ with the best and the worst estimates. **(b)** The average of all p BLFRB estimated distributions, along with the true distribution. Note that the averaged empirical distribution converges to the true cdf.

NUMERICAL EXAMPLES

Performance evaluation

The parameter settings of the BLB and BLFRB procedures for the ith element (i.e., $l = 1, \ldots, p$) of $\hat{\boldsymbol{\theta}}_n$ are as follows:

1: The bootstrap estimate of standard deviation (SD) of $\hat{\boldsymbol{\theta}}_n$ for bag k is:

$$\hat{\xi}_l^{*(k)} = \widehat{\mathrm{SD}}([\hat{\boldsymbol{\theta}}_{n,b}^{(k)}]_l) = \left(\sum_{j=1}^{r} \frac{\left([\hat{\boldsymbol{\theta}}_{n,b}^{*(kj)}]_l - [\hat{\boldsymbol{\theta}}_{n,b}^{*(k\cdot)}]_l \right)^2}{r-1} \right)^{1/2},$$

where $[\hat{\boldsymbol{\theta}}_{n,b}]_l$ denotes the lth element of $\hat{\boldsymbol{\theta}}_{n,b}$ and $[\hat{\boldsymbol{\theta}}_{n,b}^{*(k\cdot)}]_l = \frac{1}{r}\sum_{j=1}^{r} [\hat{\boldsymbol{\theta}}_{n,b}^{*(kj)}]_l$.

2: $\hat{\xi}_l^{*} = \widehat{\mathrm{SD}}([\hat{\boldsymbol{\theta}}_n]_l) = \frac{1}{s}\sum_{k=1}^{s} \widehat{\mathrm{SD}}([\hat{\boldsymbol{\theta}}_{n,b}^{(k)}]_l), \quad l = 1, \ldots, p.$

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

$$\varepsilon = \frac{\left| \widehat{\mathrm{SD}}(\hat{\boldsymbol{\theta}}_n) - \overline{\mathrm{SD}}_o(\hat{\boldsymbol{\theta}}_n) \right|}{\overline{\mathrm{SD}}_o(\hat{\boldsymbol{\theta}}_n)},$$

where $\widehat{\mathrm{SD}}(\hat{\boldsymbol{\theta}}_n) = \frac{1}{p}\sum_{l=1}^{p} \widehat{\mathrm{SD}}([\hat{\boldsymbol{\theta}}_n]_l)$ and $\overline{\mathrm{SD}}_o(\hat{\boldsymbol{\theta}}_n) = \sigma_0/\sqrt{n\mathcal{O}}$.

NUMERICAL EXAMPLES

Performance evaluation

The bootstrap setup is as follows:

- The number of distinct data subsamples (bags) is $s = 25$,
- size of each subsample is $b = \lfloor n^{\gamma} \rfloor = 1946$ with $\gamma = 0.7$ ($n = 50000$)
- maximum number of bootstrap samples in each subsample module is $r_{max} = 300$.

We start from $r = 2$ and continually add a new set of bootstrap samples (while $r < r_{max}$) to subsample modules.

NUMERICAL EXAMPLES

Performance evaluation

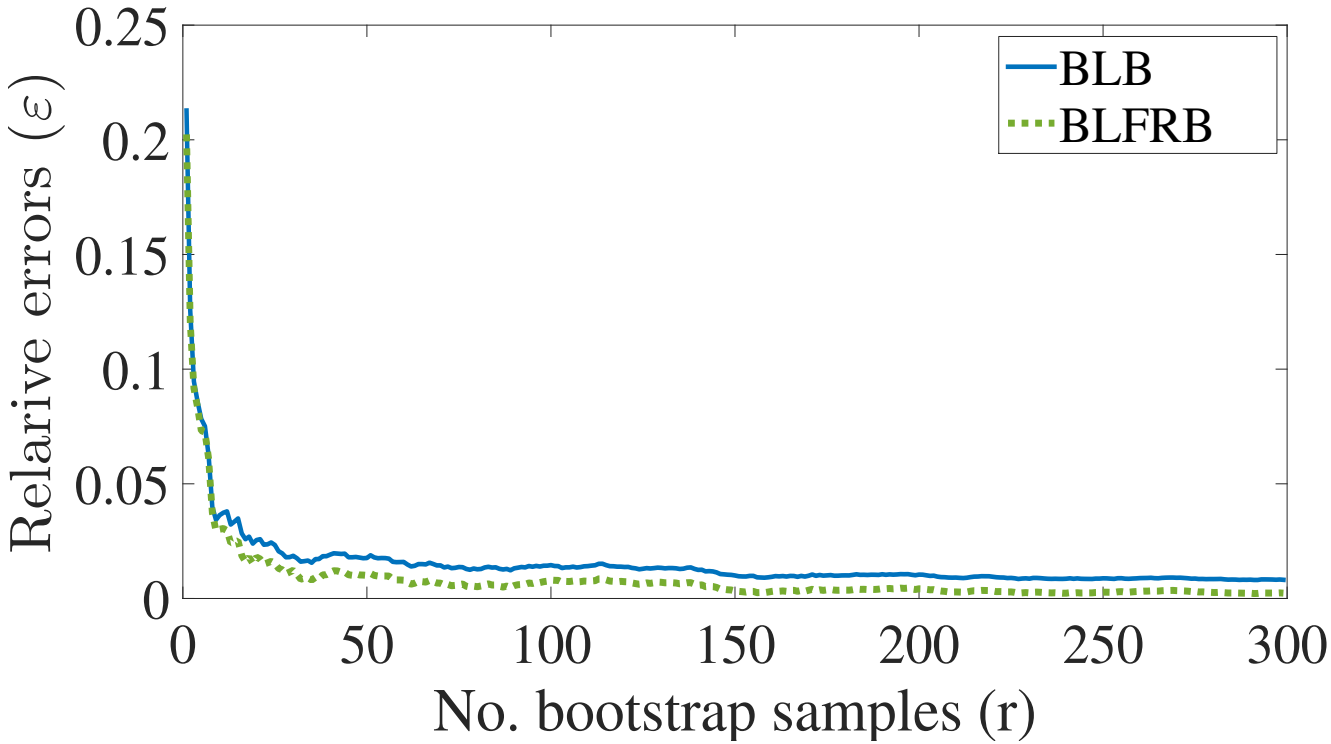
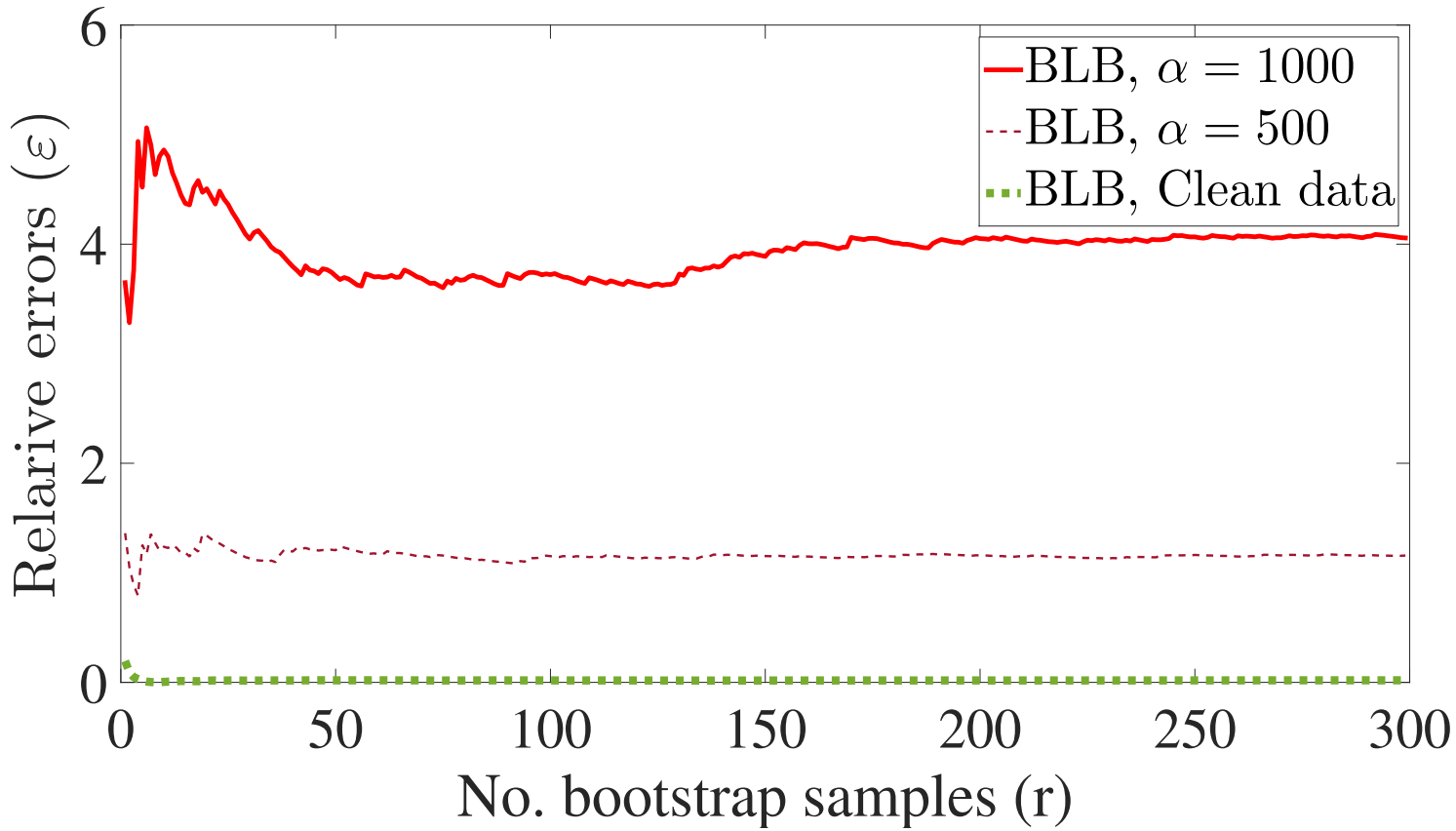


Figure: Relative errors of the BLB (solid line) and BLFRB (dashed line) methods w.r.t. the number of bootstrap samples r are illustrated. Both methods perform equally well when there are no outliers in the data.

NUMERICAL EXAMPLES

Lack of robustness of BLB

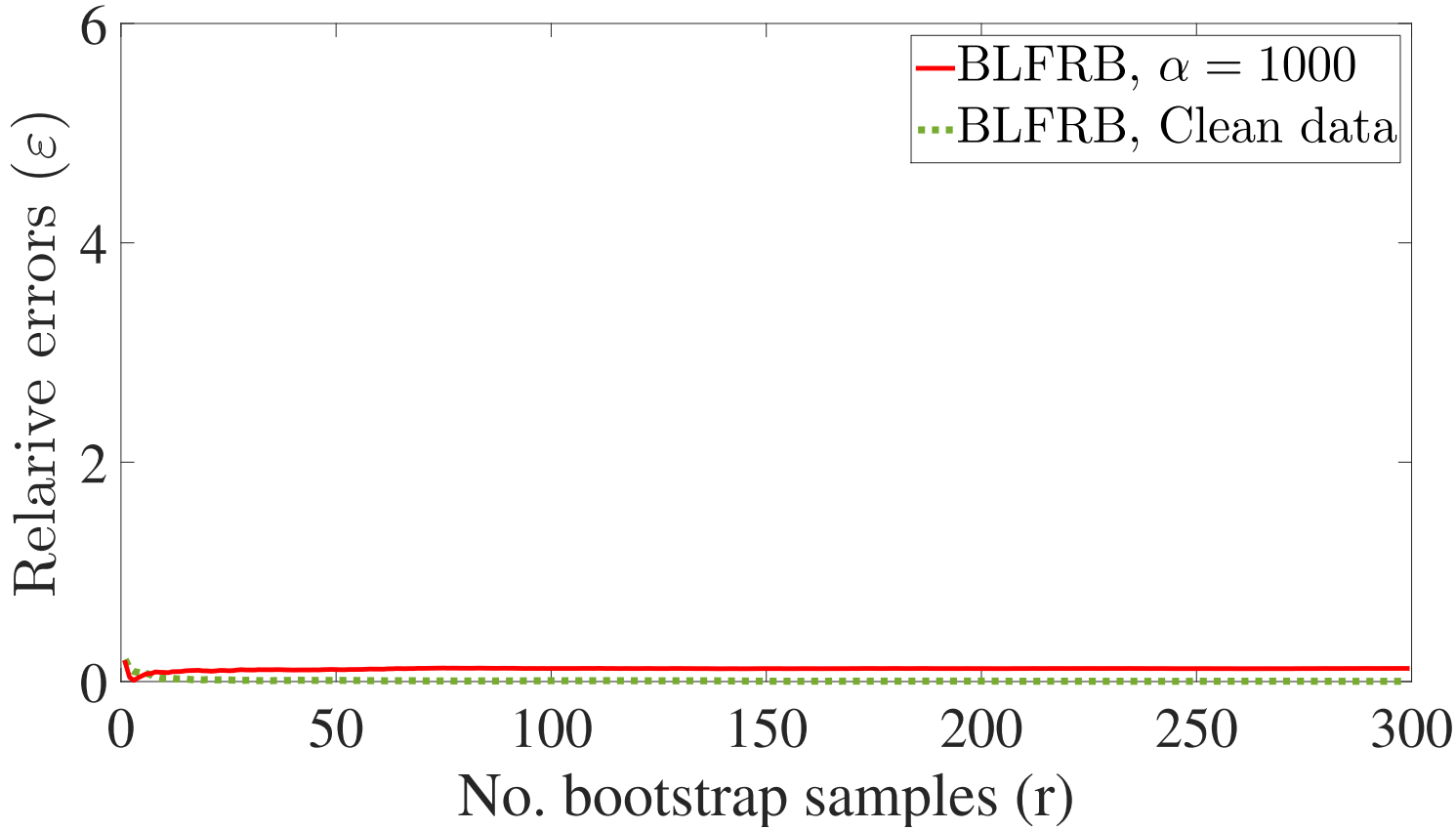
To show lack of robustness of the BLB, we introduce outlier by randomly choosing one of the data points and multiplying it by a large α.



NUMERICAL EXAMPLES

Statistical robustness of BLFRB

We severely contaminate the original data points of the first bag by multiplying 40% ($\lfloor 0.4 \times b \rfloor = 778$) of the data points by $\alpha = 1000$.



NUMERICAL EXAMPLES

Computational complexity

The computational complexity of the BLB and BLFRB methods are compared.

- The same MM-estimator is used in both schemes.
- An identical computing system is used to compute the relative errors after each iteration of the algorithms.
- The required processing time is cumulatively recorded after each iteration of the algorithms.

NUMERICAL EXAMPLES

Computational complexity

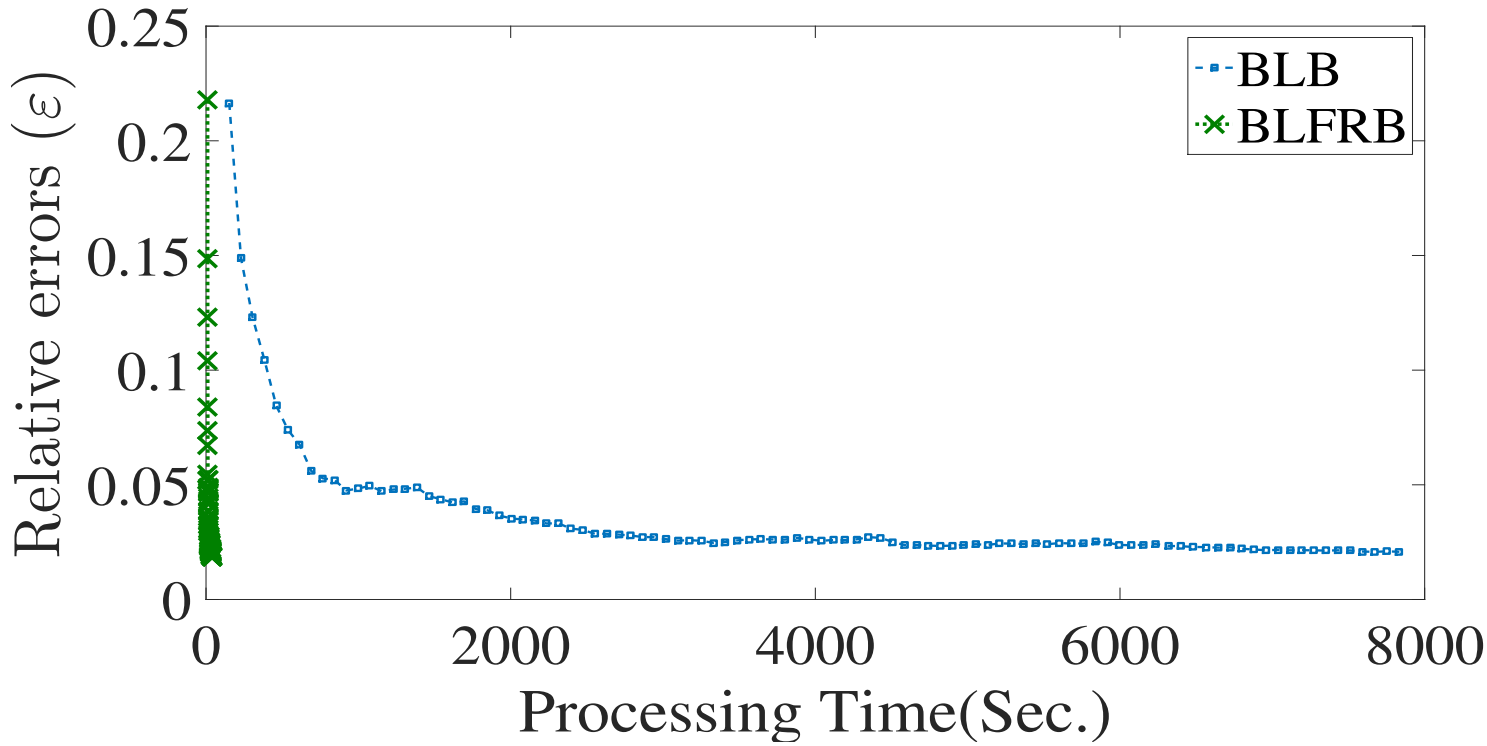


Figure: The computed relative errors ε after adding each set of new bootstrap samples are depicted w.r.t. the required processing time of computation. The BLFRB is significantly faster to compute as the (re)computation of the estimating equations is not needed in this method.

NUMERICAL EXAMPLES

Real world data

- We consider the simplified version of the the Million Song Dataset (MSD), available on the UCI Machine Learning Repository.
- The data set $\mathbf{X} = \{(y_1, \mathbf{z}_1^\top)^\top, \ldots, (y_n, \mathbf{z}_n^\top)^\top\}$ contains $n = 515345$ music tracks, where:
  - y_i (i.e., $i = 1, \ldots, n$) represents the released year of the ith song (i.e., ranging from 1922 to 2011).
  - $\mathbf{z}_i \in \mathbb{R}^p$ is a vector of $p = 90$ different audio features of each song.
  - The features are the average and non-redundant covariance values of the timbre vectors of the song.
- Linear regression can be used to predict the released year of a song based on its audio features.

NUMERICAL EXAMPLES

Real world data

- Considering the linear model $y_i = \mathbf{z}_i^\top \boldsymbol{\theta} + \sigma_0 e_i$, we use BLFRB to conduct a fast, robust and scalable hypothesis test:

$$\mathcal{H}_0 : \lfloor \boldsymbol{\theta} \rfloor_l = 0 \quad \text{vs.} \quad \mathcal{H}_1 : \lfloor \boldsymbol{\theta} \rfloor_l \neq 0,$$

  where $\lfloor \boldsymbol{\theta} \rfloor_l$ (i.e., $l = 1, \ldots, p$) denotes the lth element of $\boldsymbol{\theta}$.
- The BLFRB test of level α rejects the null hypothesis if the computed $100(1-\alpha)\%$ confidence interval does not contain 0.
- We make a test on each feature coefficient θ_i and discard it if its confidence interval contains 0

NUMERICAL EXAMPLES

Real world data

Here we run the BLFRB hypothesis test of level $\alpha = 0.05$ with the following bootstrap setup;

- Number of distinct data subsamples (bags) is $s = 51$,
- size of each subsample is $b = \lfloor n^{\gamma} \rfloor = 9964$ with $\gamma = 0.7$, $n = 515345$.
- number of bootstrap samples in each subsample module is $r = 500$.

The null hypothesis is accepted for 6 features numbered: $32, 40, 44, 47, 54, 75$. These results can be exploited to reduce the dimension of the data by excluding the ineffective variables from the regression analysis.

NUMERICAL EXAMPLES

Real world data

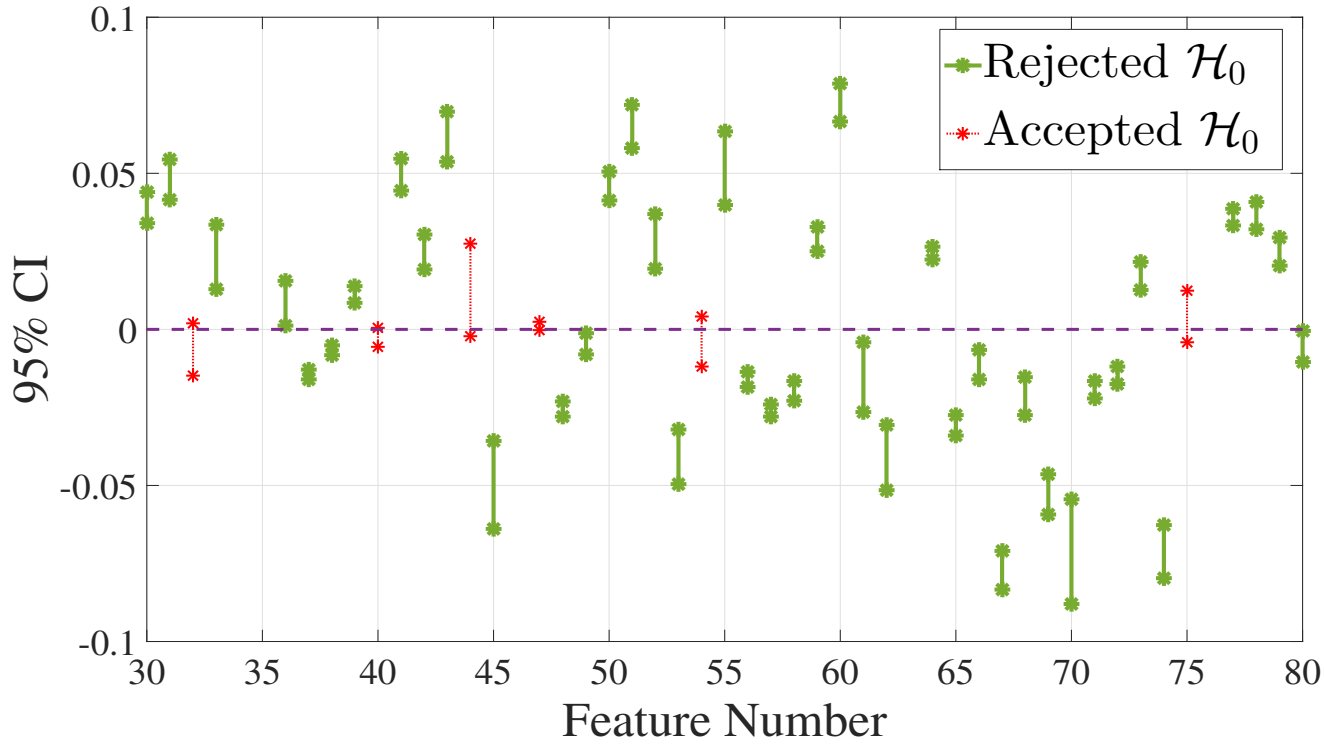


Figure: The 95% confidence intervals computed by BLFRB method is shown for some of the audio features of the MSD data set. The null hypothesis in accepted for those features having 0 inside the interval.

CONCLUSION

A new bootstrap method is introduced with the aim of facilitating bootstrap analysis of large multivariate data sets. The proposed method is:

1. Scalable to large volume and high dimensional data sets.
2. Compatible with distributed data storage systems and distributed parallel processing architectures.
3. Robust in the face of outliers.
4. Significantly faster to compute than its only counterpart (i.e., the BLB scheme).

REFERENCES

B. Efron,

Bootstrap methods: another look at the jackknife,
The Annals of Statistics, vol. 7, pp. 1-26, 1979.

P. J. Bickel, F. Gotze and W. van Zwet,

Resampling fewer than n observations: gains, losses, and remedies for losses,
Statist. Sin., vol. 7, pp. 1-31, 1997.

A. Kleiner, A. Talwalkar, P. Sarkar and M.I. Jordan.

A scalable bootstrap for massive data,
Journal of the Royal Statistical Society: Series B., 2014.

M. Salibian-Barrera, S. Van Aelst, and G. Willems,

Fast and robust bootstrap,
Statistical Methods and Applications, vol. 17, pp. 41–71, 2008.

S. Basiri, E. Ollila, and V. Koivunen,

Robust, scalable and fast bootstrap method for analyzing large scale data,
Accepted for Publication in IEEE Transactions On Signal Processing, 2015.

V. J. Yohai,

High breakdown-point and high efficiency robust estimates for regression,
The Annals of Statistics, pp. 642-656, 1987.