Recent Developments

Nooreen Dabbish, Daqian Huang, and Shinjini Nandi

April 21, 2015

Abstract

Recent Developments in Bootstrap Methodology covers parametric inference using bootstrap simulations, non-uniform nonparametric sampling, bootstrap failure, hypothesis testing, bagging, dependent data and other topics. We summarize and highlight the examples and ideas in the paper, emphasizing weighted non-parametric bootstrapping, subsampling and m out of n bootstrap, and bagging and classification. We conclude with a section on future directions suggesting what next steps can take this work further.

1 Recent Developments in Bootstrap Methodology

1.1 Introduction

This article sets out to "give a bird's eye overview of the current state of bootstrap research." The authors cover basic ideas with references to bootstrap literature as well as giving in-depth explanation and examples of extensions. Topics include parametric inference using bootstrap stimulations, non-uniform nonparametric sampling, bootstrap failure, hypothesis testing, bagging, dependent data and other topics. Our goal is to summarize and highlight the examples and ideas in the paper. We place special emphasis on weighted non-parametric bootstrapping, subsampling and m out of n, and bagging and classification. We conclude with a section on future directions suggesting what next steps can take this work further.

1.2 Basic Ideas: Bootstrap approaches to confidence intervals and hypothesis testing

In an overview of Bootstrap approaches to confidence intervals and hypothesis testings, the authors describe how non parametric CIs are obtained, either through Studentized pivots or the direct use of bootstrap quantiles. Studentized pivots that use the bootstrap estimated variance, $V^\star$ can be constructed. An Edgeworth expansion (similar to a Taylor series, but of a probability density function) is used to demonstrate the bootstrapped pivots are consistent estimators for "smooth" estimators. They are accurate to 1-α + $O(n^{-1})$, an improvement of $O(n^{-1/2})$.

The authors explain that using a pivot is beneficial because it avoids the need to modify the samping plan in hypothesis testing. They further describe model-based bootstrapping when data are not identically distributed for example time-series/autoregressive moving average.

1.3 Bootstraps for Parametric Likelihood Inference

To show bootstrap's use for parametric likelihood inference, the authors compare confidence set coverage calculations for an exponential regression problem. They use the N(0,1) distribution of the signed root likelihood ratio statistic, r_p, and find that it is not as accurate as the N(0,1) distributed of the adjusted signed root r_a or parametric bootstrap confidence sets. While the adjusted r_a calculation requires the use of an ancillary statistic, the authors point out that the bootstrap does not necessitate this additional analytic step.

In a second example of normal distributions with common mean, different variances, r_a is intractable. The authors show that for a six sample data set of cotton yarn, it possible to use MLE or Constrained MLE bootstrap to attain near-nominal coverages for confidence sets.

1.4 Weighted Non-parametric Bootstrapping

1. Why do we need weighted nonparametric bootstrapping?
   As we all know, the ordinary nonparametric bootstrap uses uniform resampling from original data sample to simulate bootstrap sample. A merit of these way is that we can obtain a generally reliable nonparametric confidence intervals. There are two approaches. The first method is called studentized bootstrap. This is inspired by the student t statistic and requires an estimate variance V^* for $\hat{\theta}^*$ based on the same bootstrap sample. Then using Edgeworth expansion, a wide class of estimators $\hat{\theta}$ will be derived from the quantiles of $Z^* = (\hat{\theta}^* - \hat{\theta})/V^{*1/2}$. The second approach is that resampling of $\hat{\theta}^*$ conditional on $\hat{\theta}$ is used to approximate sampling from the posterior distribution of θ given $\hat{\theta}$. This interval is known as bisas-corrected and accelerated (BC_a) interval.
   But numerical work has shown that both studentized bootstrap and BC_a intervals typically show slight undercoverage since occasional instability in the variance estimate V can lead to excessively long intervals.
   In order to avoiding this undercoverage, the process of prepivoting was come out by Beran (1987, 1988).
   The main idea is that resampling from a nonuniform distribution $\widetilde{F}_0$ with the constraint that $\theta(\widetilde{F}_0) = \theta_0$.

2. How does weighted nonparametric bootstrapping work?
   Suppose that there is an unknown distribution F and one parameter of interest θ. We want to estimate F nonparametrically under the constraint $\theta(F) = \theta_0$, where θ_0 may not be the true value of θ. Y_i is the original data coming from F. Then given θ_0 and a data set Y, we use arbitrary probability $p = (p_1, p_2, ..., p_n)$, where $\sum_i^n p_i = 1$, to weight Y_i. Next, we choose $p = p(\theta_0)$ to minimize the Kullback-Leibler distance between $\hat{F}_p$ and $\hat{F}$,

$$\int log \frac{d\hat{F}_p}{d\hat{F}} d\widetilde{F}(x) = -\frac{1}{n} \sum_{j=1}^{n} log(np_j)$$

   with constraint $\theta(\hat{F}_p) = \theta_0$.
   Here is a theorem [5] for solving this $p_i(\theta)$.
   Theorem 1 For $\mu \in (y_{(1)}, y_{(n)})$,

$$p_i(u) = \frac{1}{n - \lambda(y_i - \mu)} > 0, \qquad 1 \le i \le n,$$

   where λ is the unique solution of the equation

$$\sum_{j=1}^{n} \frac{y_j - \mu}{n - \lambda(x_j - \mu)} = 0$$

in the interval $(\frac{n}{x_{(1)}-\mu}, \frac{n}{x_{(n)}-\mu})$.

When we get p_i, we can use this $\hat{F}_p$ as resampling distribution to do weighted bootstrap. Let $Y^\dagger$ be the bootstrap sample from above nonuniform distribution $\hat{F}_p$.

By using prepivoting method, we construct an approximately uniform random variable $U = u(Y, \theta)$, a transforming function on (0, 1). Such that a one-side confidence set for θ is $\{\theta : u(Y, \theta) \leq 1-\alpha\}$. According to the percentile method, $u(Y, \theta) = Pr^*(\hat{\theta}^* > \theta)$, where the asterisk indicates uniform bootstrapping from Y. On the other hand, based on normal approximation, a confidence set of asymptotic coverage $1-\alpha$ can be defined by $u(Y, \theta) = \Phi\{(\hat{\theta} - \theta)/V^{1/2}\}$.

The uniform bootstrap estimates the distribution function $G(x|\theta)$ of $u(Y, \theta)$ by

$$\hat{G}(x) = Pr^*\{u(Y^*, \theta) \leq x\}$$

from which we can define the prepivoted root $\hat{u}_1(Y, \theta) = \hat{G}\{u(Y, \theta)\}$.
Similarly, the weighted bootstrap estimates the distribution function by

$$\hat{G}(x) = Pr^\dagger\{u(Y^\dagger, \theta) \leq x\}$$

from which we can define the prepivoted root $\widetilde{u}_1(Y, \theta) = \widetilde{G}\{u(Y^\dagger, \theta)\}$.

Method	formula	reduced error by
uniform	$Pr\{u(Y, \theta) \leq \mu\} = \mu + O(n^{-j/2})$	1
uniform bootstrap	$Pr\{u(Y^*, \theta) \leq \mu\} = \mu + O(n^{-(j+1)/2})$	$O(n^{-1/2})$
weighted bootstrap	$Pr\{u(Y^\dagger, \theta) \leq \mu\} = \mu + O(n^{-(j+2)/2})$	$O(n^{-1})$

Lee and Young showed that this conclusion applies to regression settings and robust inference, as well as to more conventional problems within the smooth function model. Compared to conventional bootstrapping, weighted bootstrap prepivoting accelerates the rate of convergence of the error of the bootstrap inference.

3. Expectation
   Even though weighted bootstrap have above merits, it still not easily solve a set of parameters or null hypotheses.

1.5 Subsampling and the m out of n bootstrap

Definition of Subsampling: Suppose we have a sample

$$\mathbf{Y} = Y_1, \ldots, Y_n \text{ from a distribution } F$$

Subsampling is drawing smaller samples or subsamples of size $m < n$ from $\mathbf{Y}$ without replacement in order to draw statistical conclusions about any parameter of interest.

The required statistic is calculated from each sample and these values are used to construct an approximation of the appropriate sampling distribution.

Advantages of Subsampling over ordinary bootstrap

Subsampling achieves more accuracy in estimation than ordinary bootstrap under less stringent regularity conditions than those required by bootstrap.

It provides asymptotic consistency even under extreme weak conditions and is hence a popular non-parametric tool.

Some of its advantages over bootstrap are given below.

1. Subsampling has a distinct advantage over bootstrapping when the sample available is from time series data. Time series is dependent data, i.e., the ordering of data according to time is extremely important. It makes little or no sense in drawing identical and independent resamples using bootstrap technique from time series data. Using sub-sampling technique, we can obtain samples and yet maintain the dependence according to time. The statistics obtained may not be the unbiased estimates required, but they can be easily amended to obtain the required estimates. For example, if $(Y_1, \ldots, Y_n)$ is a time series, the subsequences $(Y_s, \ldots, Y_{s+m-1}),\ s = 1, \ldots, S,\ S = n - m + 1$ are the subsamples.

2. When constructing confidence region for parameters. Suppose we have sample

$$\mathbf{Y} = Y_1, \ldots, Y_n \text{ from an unknown distribution } F$$

and $\theta \equiv \theta(F)$ is the parameter for which we want a confidence region. The confidence region is constructed from a statistic $\hat{\theta}_n$ that converges weakly to θ at a rate τ_n.(A common choice of τ_n is $\sqrt{n}$). Let $\hat{\sigma}_n$ be an estimator of σ, σ^2 being the asymptotic variance of $\tau_n\hat{\theta}_n$. Suppose $J_n(F)$ be the sampling distribution of $\frac{\tau_n(\hat{\theta}_n-\theta)}{\hat{\sigma}_n}$. Let $J(F)$ be a non-degenerate distribution such that

$$J_n(x,\, F) \to J(x,\, F)\ as\ n \to \infty$$

We draw $S = \binom{n}{m}$ sub-samples each of size m and let $\hat{\theta}_{n,m,s}$ and $\hat{\sigma}_{n,m,s}$ be the estimates of $\hat{\theta}$ and $\hat{\sigma}$ obtained from the \$s\$th sub-sample. The sub-sampling distribution of $\frac{\tau_n(\hat{\theta}_n-\theta)}{\hat{\sigma}_n}$ is

$$L_{n,m}(x) = \frac{\sum_{s=1}^{S} I[\frac{\tau_n(\hat{\theta}_n-\theta)}{\hat{\sigma}_n} \le x]}{S}$$

If $m \to \infty$ and $max(\frac{m}{n},\ \frac{\tau_m}{\tau_n}) \to \infty$ as $n \to \infty$ then

$$sup_x[L_{n,m}(x) - J(x, F)] = o_p(1)$$

Hence quantiles of $L_{n,m}$ can be used to construct confidence sets of θ with asymptotically correct coverage. Bootstrap approximations can provide analogous results, but under more strong assumptions. For example, the convergence of $J_n(F)$ to $J(F)$ must be locally uniform in F . Nonuniformity in convergence is responsible for bootstrap failure.

3. Subsampling can be used to remedy bootstrap. If subsampling is done with replacement, we get *m out of n* bootstrap .
   With assumption $m/n \to 0$ and additional assumption $m^2/n \to 0$, sub-sampling and *m out of n* bootstrap give similar asymptotically valid conclusions.

Disadvantages of Subsampling

1. The proper choice of the sub-sample size is difficult. If the choice is not perfect, high-order accuracy cannot be obtained from sub-sampling.
2. When bootstrap is valid, it is usually preferred to sub-sampling. When *m out of n* bootstrap is valid, it is also preferred to sub-sampling.

But if validity of bootstrap cannot be verified or under minimal conditions or complicated data structures, sub-sampling is always the preferred solution.

1.6 Bootstrapping Superefficient estimators

The authors example a case of bootstrap failure. They cite historical work showing that the consistency of conventional bootstrap depends on true value of θ (Beran 1997). The Stein estimator, prototypical of many nonparametric smoothers, for $Y_1, \ldots, Y_n \overset{iid}{\sim} N_k(\theta, I)$ is given by $T = \left(1 - \frac{k-2}{n||Y||^2}\right) \bar{Y}$.

They argue the bootstrap estimator $\mathrm{H}(\cdot, \bar{Y})$ is consistent for $\theta \neq 0$, but inconsistent when $\theta = 0$. They show simulation results for coverages, including a modified bootstrap where the estimator $\hat{\theta}$ is set to $\bar{Y}$ above a threshold and 0 below. The modified parametric results at $\theta = 0$ are good, with a steep price in loss of accuracy (overly conservative sets) for small θ. Also, the conventional parametric bootstrap constructs sets with higher-than-nomial coverage near 0.

1.7 More on Significance Tests

Next, the authors underscore bootstrap's role in comparing nonparametric model fit tests to parametric model fits or fits from several data sets. They define ESP, the Empirical Strength Probability. Given $H_0 : \theta \in \Theta_0$, ESP = proportion of $\tilde{\theta}_r^\star \in \Theta_0$. ESP acts like a

p-value asymptotically as $n \to \infty$. In one ESP example, the authors look at the exponential mean with Y_i independent $\sim \exp(\mu\}$, $H_0 : \mu \leq \mu_o, H_1 : \mu > \mu_0$. Here, an exact test has a $p \sim U$. Parametric bootstrap ESP does not work well for small sample, nonparametric works well. Additionally, to test the fit of a specific distribution, define distance between distributions and use d($\tilde{F}$,F_0)

1.8 Bagging and Classification

Introduced by Breiman 1996 as a way of improving unstable classification and prediction algorithms, "bagging" is boostrap aggregation. A data set, often called the learning set or training set, is resampled with replacement to create R learning sets. From these, R classification or prediction schemes are generated. Their predictions are averaged to create the bagged predictor for a continuous numerical response. For a class response, the bagged predictor selects the class that received the highest number of votes.

Buhlmann and Yu (2002), recent work highlighted here used theory and simulation to show that bagging functions to reduce variance by converting certain hard thresholding predictors to soft thresholding. They build their case starting with the example of a bagged indicator function. The indicator initially has a step shape, and bagging converts it to a sigmoidal, inverse probit shape. In this case, they show that the bagged function is asymptotically unbiased, but has a reduction in asymptotic variance by 1/3. Building on this example, Buhlmann and Yu show that when screening predictor variables in linear regression formula, bagging can reduce MSE for the predictor by up to 50%.

An additional recent development underscored by Davidson, Hinkley and Young is "boosting." In boosting, data that is difficult to classify is typically (but not always) given greater weight so that future learners focus more on previously misclassified data. Boosting can improve on classification error, even with respect to bagged classifiers.

1.9 Bootstrapping Dependent Data

In a discussion on the use of bootstrap for dependent data, such as time series, stochastic processes, and spatial data, the authors discuss block sampling. They point out that the type of spatial data in question may determine the appropriate sampling technique ans suggest that the area is ripe for future work. Finally, they mention model selection, hierarchical and random effects models and point out hierarchical model bootstrapping is "underdeveloped."

2 Future Directions

Based on the authors suggestions, two hot areas for research are bootstrapping dependent data and the nonparametric or semi-parametric hierarchical model bootstrap. We find

it interesting that these areas are seemingly connected through their inherent and rich covariance structure and believe they are linked theoretically as well.