Bootstrap Methods

Erniel B. Barrios, Ph.D.
University of the Philippines Diliman

The classical framework of statistical inference relies heavily on the sampling distribution as a link between the information provided by the sample and the generalizations it provides about the population. Samples are drawn independently and assumed to come from some special distributions so that a closed-form sampling distribution is easily derived. In the more complicated scenario, desirable properties are derived invoking large sample sizes in cases where the sampling distribution is not mathematically tractable. Many statistics are usually analyzed dealing with small samples, often resulting to more complicated standard errors.

Developments in statistical inference had been influenced tremendously by access to efficient computing facilities that allow verification of properties of complicated statistics or those without closed-form. The Bootstrap is a resampling method involving large amount of computations that facilitates small sample inference on a variety of estimation and hypothesis testing problems.

Efron (1979) introduced the bootstrap as a special case of the jackknife, a resampling method already known much earlier. The method was originally intended for independent set of observations, say a random sample $x = (x_1, x_2, ...x_n)$ from F. With the aim of understanding the sampling distribution of the random variable $R(x, F)$ from *x*, Efron (1979) provided the following bootstrap algorithm:

1. Construct the sample probability distribution $\hat{F}$, putting mass $1/n$ at each of the mass point $x_1, x_2, ..., x_n$.
2. With $\hat{F}$ fixed, draw random sample of size n from $\hat{F}$, say $X_i^* = x_i^*$, $X_i^* \sim iid\ \hat{F}$, i = 1,2,...,n with replacement. The bootstrap sample is composed of the set $x^* = (x_1^*, x_2^*, ..., x_n^*)$. m replicates of *x** is generated where m is reasonably large.
3. Approximate the sampling distribution of by the distribution of computed from the m replicates.

The bootstrap estimate of the mean and variance of the sampling distribution of R are:

$$R_B = \frac{\sum_{j=1}^{m} R_j^*}{m} \qquad V(R_B) = \frac{\sum_{j=1}^{m} (R_j^* - R_B)^2}{m}$$

Resampling is intended to smoothen the estimate by trimming off the bias that the nuisance of the sample selection might create. The method was illustrated to be applicable in variance estimation for statistics with complicated sampling distributions and for modeling problems like regression analysis.

Starting from the computationally attractive value of the bootstrap method, analytical properties were established, e.g., consistency provided by the Glivenko-Cantelli Theorem [Given $x_1, x_2, ..., x_n$ *iid* F let $F_n(x) = \frac{1}{n}\sum_{m=1}^{n} I_{(x_m \le x)}$. As $n \to \infty$ $\sup_x |F_n(x) - F(x)| \to 0$ a.s.]. Davison and Hinkley (1997: 31-38) observed that error in the bootstrap are classified into statistical error [the small difference between the true distribution *F* and the estimate $F_n(x)$] and the simulation error [since properties of statistics are approximated by empirical properties in simulation, influenced by factors like resample size, replication size, etc.]. Resampling size is an important factor that could influence the error in the bootstrap, for example, (Bickel et al., 1997) resampled with m<n observations and concluded that this approach can be expected to exhibit relative advantage on small samples.

The theoretical basis of the bootstrap has continued to be provided by over 1,000 papers since 1979 (Efron, 2000). While the bootstrap continued to define the landscape on the interplay between computing and statistical inference, there are also reminders that this is not the ultimate solution to all statistical problems. Beran (1997) recognized the viability of convergence of the bootstrap to the correct limiting distribution, but noted that convergence fails at superefficiency points in the parameter space. Furthermore, superefficiency is only a sufficient condition for bootstrap failure. Andrews (2000) further cautioned the bootstrap is not a universal solution to statistical inference problems, also provided counterexample illustrating that bootstrap is inconsistent when the parameter is on the boundary of the parameter space.

Efron also introduced the bootstrap in the context of model-based inference, where instead of the random sample, resampling is performed on the empirical distribution of the residuals. For the regression model $X_i = g_i(\beta) + \varepsilon_i$ i=1,2,..., n, $\varepsilon_i \sim$ *iid* F the parameter β is estimates by $\hat{\beta}$ (e.g., ordinary least squares). The sampling

distribution is defined as: $\hat{F}: mass\ \frac{1}{n}$ at $\hat{\varepsilon}_i = x_i - g_i(\hat{\beta}),\ i = 1,2,...,n$. From the pair $(\hat{\beta}, \hat{F})$, the bootstrap sample is computed from the fitted model as $X_i^* = g_i(\hat{\beta}) + \varepsilon_i^*,\ \ \varepsilon_i^* \sim iid\ \ \hat{F}$. Each of the bootstrap samples can provide an estimate of β following the same estimation procedure used (e.g., ordinary least squares). From all the bootstrap replicates, we get $\hat{\beta}*^1, \hat{\beta}*^2, ..., \hat{\beta}*^m$ and used to estimate the distribution of $\hat{\beta}*$. In model-based inference, Paparoditis and Politis (2005) underscored the importance of the choice of residuals. For example, to maximize power in bootstrap-based hypothesis testing, residuals are obtained using a sequence of parameter estimators that converge to the true parameter value both under the null and alternative hypothesis.

The bootstrap was initially introduced for independent cross-section data, but recently, it has been defined for time series data and other dependent observations as well. There are many theoretical justifications of time series bootstrap, example, Politis and Romano (1994) established convergence of certain sums of stationary time series that can facilitate bootstrap resampling. The block bootstrap was among the early proposal for time series data. While the method is very straightforward, there are associated problems like independence of block to maintain the dependence structure within the block. The size of the block is a crucial quantity that should be determined to assure success in block bootstrap. The AR-sieve was also introduced as a residual-based method similar to the model-based approach. Local bootstrap was also introduced but in the context of local regression framework (nonparametric) and to account for the nonparametric model, resampling allows the empirical distribution to vary locally in the time series. Bühlman (2002) compared different methods for time series bootstrap. The block bootstrap is recognized as the most general and simple generalization of the original independent resamples but is criticized for the possible artifacts it may exhibit when blocks are linked together. Blocking can potentially introduce some dependence structure in addition to those naturally existing in the data. The AR-sieve is less sensitive to selection of a model than the block length. The local bootstrap for nonparametric estimation is observed to yield slower rate of convergence. Generally, the AR-sieve is advantageous among the bootstrap approaches for time series data.

Recently, the bootstrap has been introduced to more complex situations and in more complicated models. In modeling nonstationary volatility, Xu (2008) used autoregression around a polynomial trend with stable autoregressive roots to illustrate how nonstationary volatility affects the consistency, convergence rates and asymptotic distributions of the estimators. Westerlund and Edgerton (2007) proposed a bootstrap test for the null hypothesis of cointegration in panel data. Dumanjug, et al. (2010)

developed a block bootstrap method in a spatial-temporal model. Chernick, et al. (2010) also provide a comprehensive summary of the development of the bootstrap method as it identifies a good range of literature on the subject matter.

REFERENCES

ANDREWS, D., 2000, Inconsistency of the Bootstrap When a Parameter is on the Boundary of the Parameter Space, *Econometrica*, 68(2): 399-405.

BERAN, R., 1997, Diagnosing Bootstrap Success, *Annals of the Institute of Statistical Mathematics*, 49(1): 1-24.

BICKEL, P., GÖTZE, F., AND VAN ZWET, W., 1997, Resampling Fewer Than Observations: Gains, Losses, and Remedies for Losses, *Statistica Sinica*, 7: 1-31.

BÜHLMAN, P., 1997, Sieve Bootstrap for Time Series, *Bernoulli*, 3(2): 123-148.

BÜHLMAN, P., 2002, Bootstrap for Time Series, S*tatistical Science*, 17(1): 52-72.

CHERNICK, M., GONZALEZ-MANTEIGA, W., CRUJEIRAS, R., AND BARRIOS, E., 2010, Bootstrap Methods, in M. Lovric, ed., *International Encyclopedia of Statistical Sciences*, N.J: Springer.

DAVISON, A. AND HINKLEY, D., 1997, *Bootstrap Methods and Their Applications*, Cambridge:Cambridge University Press.

DUMANJUG, C., BARRIOS, E., AND LANSANGAN, R., 2010, Bootstrap Procedures in a Spatial-Temporal Model, *Journal of Statistical Computing and Simulation*, 80(7):809-822.

EFRON, B., 1979, Bootstrap Methods: Another Look at the Jackknife, *The Annals of Statistics*, 7(1): 1-26.

EFRON, B., 2000, The Bootstrap and Modern Statistics, *Journal of the American Statistical Association*, 95(452): 1293-1296.

PAPARODITIS, E., AND POLITIS, D., 2005, Bootstrap Hypothesis Testing in Regression Models, *Statistics and Probability Letters*, 74: 356-365.

POLITIS, D., AND ROMANO, J., 1994, Limit Theorems for Weakly Dependent Hilbert Space Valued Random Variables with Application to the Stationary Bootstrap, *Statistica Sinica*, 4: 461-476.

WESTERLUND, J., AND EDGERTON, D., 2007, A Panel Bootstrap Cointegration Test, *Economic Letters*, 97: 185-190.

XU, K., 2008, Bootstrapping Autoregression Under Nonstationary Volatility, *Econometrics Journal*, 11: 1-26.