Resampling Fewer Than n Observations: Gains, Losses, and Remedies for Losses

By

P.J. Bickel[1]
Dept. of Statistics, UC Berkeley

F. Götze
Fakultät für Mathematik, Universität Bielefeld

W.R. van Zwet[1]
Dept. of Mathematics, University of Leiden

[1]Research supported by SFB 343 "Diskrete Strukturen in der Mathematik", Universität Bielefeld

1. Introduction

Over the last 10-15 years Efron's nonparametric bootstrap has become a general tool for setting confidence regions, prediction, estimating misclassification probabilities, and other standard exercises of inference when the methodology is complex. Its theoretical justification is based largely on asymptotic arguments for its consistency or optimality. A number of examples have been addressed over the years in which the bootstrap fails asymptotically. Practical anecdotal experience seems to support theory in the sense that the bootstrap generally gives reasonable answers but can bomb.

In a recent paper Politis and Romano (1993), following Wu (1990), and independently Götze (1993) showed that what we call the m out of n without replacement bootstrap with $m = o(n)$ works to first order both in the situations where the bootstrap works and those known so far where it does not. In this paper we

- Investigate this phenomenon further and show the extent to which it is shared by the m out of n with replacement bootstrap which we show has some advantages. The m out of n with replacement bootstrap with $m = o(n)$ has been known to work in most of the examples of bootstrap failure.

- Study the price in efficiency when the usual bootstrap works that is paid for protection in case it doesn't and show how, by the use of extrapolation the price can be avoided.

- Support some of our theory with simulations.

With the exception of theorem 2 and corollary 1 we do not present essentially new results.

We suppose throughout that we observe $X_1, \ldots, X_n$ taking values in $\boldsymbol{X} = R^p$ (or more generally a separable metric space.) i.i.d. according to $F \in \mathcal{F}_0$. We are interested in the distribution of a symmetric function of $X_1, \ldots, X_n$; $T_n(X_1, \ldots, X_n, F) \equiv T_n(\hat{F}_n, F)$ where $\hat{F}_n$ is defined to be the empirical distribution of the data. More specifically we wish to estimate a parameter which we denote $\theta_n(F)$, of the distribution of $T_n(\hat{F}_n, F)$, which we denote by $\mathcal{L}_n(F)$. We will usually think of θ_n as real valued, for instance, the variance of $\sqrt{n}$ median $(X_1, \ldots, X_n)$ or the 95% quantile of the distribution of $\sqrt{n}(\bar{X} - E_F(X_1))$.

Suppose θ_n is defined naturally not just on $\mathcal{F}_0$ but on $\mathcal{F}$ which is large enough to contain all discrete distributions. It is then natural to estimate F by the nonparametric maximum likelihood estimate, $\hat{F}_n$, and hence $\theta_n(F)$ by $\theta_n(\hat{F}_n)$. This is Efron's (ideal) nonparametric (np) bootstrap. Since $\theta_n(F) \equiv \gamma(\mathcal{L}_n(F))$ and, in the cases we consider, computation of γ is straightforward the real issue is estimation of $\mathcal{L}_n(F)$. Efron's (ideal) bootstrap is to estimate $\mathcal{L}_n(F)$ by the distribution of $T_n(X_1^*, \ldots, X_n^*, \hat{F}_n)$ where, given $X_1, \ldots, X_n$, the X_i are i.i.d. $\hat{F}_n$, i.e. the bootstrap distribution of T_n. In practice, the bootstrap distribution is itself estimated by Monte Carlo or more sophisticated resampling schemes, (see deCiccio and Romano (1988) and Hinkley (1988)). We will not enter into this question further.

Theoretical analyses of the bootstrap and its properties necessarily rely on asymptotic theory, as $n \to \infty$ coupled with simulations. We follow this path and impose a minimum