ADAPTIVE CHOICE OF BOOTSTRAP SAMPLE SIZES

FRIEDRICH GÖTZE[1] AND ALFREDAS RAČKAUSKAS[2]

University of Bielefeld and University of Vilnius

Consider sequences of statistics $T_n(\widehat{P}_n, P)$ of a sample of size n and the underlying distribution. We analyze a simple data–based procedure proposed by Bickel, Götze and van Zwet (personal communication) to select the sample size $m = m_n \leq n$ for the bootstrap sample of type "m out of n" such that the bootstrap sequence T_m^* for these statistics is consistent and the error is comparable to the minimal error in that selection knowing the distribution P. The procedure is based on minimizing the distance between $L_m(\widehat{P}_n)$ and $L_{m/2}(\widehat{P}_n)$, where $L_m(\widehat{P}_n)$ denotes the distribution of T_m^*.

AMS subject classifications: 62D05.
Keywords and phrases: Bootstrap, m out of n bootstrap, Edgeworth expansions, model selection.

1 Introduction

In this paper, we investigate an adaptive choice of the bootstrap sample size m in sampling from an i.i.d. sample of size n m-times independently and with (resp., without) replacement. To simplify the writing we shall abbreviate the notion of *m out of n* sampling as *moon bootstrap.*

Assume that the random elements $X_1, \ldots, X_n, \ldots$ are independent and identically distributed from a distribution P on a measurable space $(S, \mathcal{A})$. Let $\widehat{P}_n$ denote the empirical measure of the first n observations $X_1, \ldots, X_n$. Throughout we assume that $P \in \mathcal{P}_o \subset \mathcal{P}$, where $\mathcal{P}_o$ is a set of probability measures on $(S, \mathcal{A})$ containing all empirical measures $\widehat{P}_n$.

Let $T_n = T_n(X_1, \ldots, X_n; P)$ denote a sequence of statistics, possibly dependent on the unknown distribution P in order to ensure that T_n is weakly convergent to some limiting distribution as n tends to infinity. A typical example is given by $T_n = n^\alpha(F(\widehat{P}_n) - F(P))$, where $F : \mathcal{P}_0 \to R$ denotes a functional on $\mathcal{P}_0$ and $\alpha > 0$ is an appropriate normalization rate.

We are interested in the estimation of the distribution function (d.f.) $L_n(P; a)$ of T_n by means of resampling methods.

[1] This research was supported by the Deutsche Forshungsgemeinshaft SFB 343.

[2] This research was supported by the Deutsche Forshungsgemeinshaft SFB 343 and Institute of Mathematics and Informatics, Lithuania.