

10 Statistics

10.10 Moving Blocks Bootstrap

(5 units)

This project requires an understanding of the Part IB course Statistics. Bootstrap methods are covered in more detail in the Principles of Statistics course.

1 Introduction

Bootstrap methods are procedures for the empirical estimation or approximation of sampling distributions and their characteristics. Their primary use lies in the estimation of accuracy measures, such as bias and variance, for parameter estimators, and in construction of confidence sets or hypothesis tests for population parameters. They are applied in circumstances where the form of the population from which the observed data has been drawn is unknown.

The bootstrap principle was formalized by Efron [1]. It may be summarized for a general situation as follows. We have data $Y = (Y_1, \dots, Y_n)$ (not necessarily independent and identically distributed) and a statistical model P under which the data are obtained. Usually, P can be described by the joint distribution of Y , or by some quantities that uniquely determine this joint distribution. Suppose we wish to estimate the distribution of a random variable or ‘pivot’ $R_n(Y; P)$, or some characteristic of that distribution. Then the data Y is used to estimate P by \hat{P} . Letting Y^* be a bootstrap data set generated from \hat{P} , then the bootstrap estimator of the distribution of $R_n(Y; P)$ is the conditional distribution of $R_n(Y^*; \hat{P})$, given Y . Where this conditional distribution is not expressible as an explicit function of Y , Monte Carlo simulation can be used to construct an approximation to the bootstrap estimator.

Bootstrap methods are most fully developed for the case where Y_1, \dots, Y_n are an independent and identically distributed (IID) sample. The purpose of this project is to investigate a way of extending the bootstrap idea to dependent data.

2 Moving blocks bootstrap

Consider estimation of the sampling distribution of $\bar{Y} = n^{-1} \sum_{i=1}^n Y_i$, but suppose that the Y_i are m -dependent. We define a sequence of random variables $\{Y_n, n = 0, \pm 1, \pm 2, \dots\}$ to be **stationary** if, for all $n = 0, \pm 1, \pm 2, \dots$, the joint distribution of $(Y_k, Y_{k+1}, \dots, Y_{k+n})$ does not depend on k , and we define a stationary sequence to be **m -dependent** if the set of random variables $\{Y_n, n = -1, -2, \dots\}$ is independent of $\{Y_n, n = m, m + 1, \dots\}$.

If Y_1, \dots, Y_n are from a univariate m -dependent stationary sequence, $E(\bar{Y}_n) = \mu$, with $\mu = E(Y_1)$, and if we define $\sigma_n^2 = \text{var}(\sqrt{n}\bar{Y}_n)$ then we see that, for $m \leq n$,

$$\sigma_n^2 = \text{var}(Y_1) + 2 \sum_{i=1}^m \left(1 - \frac{i}{n}\right) \text{cov}(Y_1, Y_{1+i}).$$

You may assume that by the Central Limit Theorem for m -dependent processes, $\sqrt{n}(\bar{Y}_n - \mu)$ converges in distribution to $N(0, \sigma_\infty^2)$, where

$$\sigma_\infty^2 = \lim_{n \rightarrow \infty} \sigma_n^2 = \text{var}(Y_1) + 2 \sum_{i=1}^m \text{cov}(Y_1, Y_{1+i}).$$

Künsch [2] proposed a “moving blocks” resampling scheme for stationary time series data. The basic idea is to break the observed data series Y up into a collection of overlapping blocks of observations. Bootstrapped data series are obtained by independent sampling, with replacement, from among these blocks.

We consider this procedure in the context of the example above. Let the given data be Y_1, \dots, Y_n and b be a given block size. Define $\xi_i = (Y_i, \dots, Y_{i+b-1})$ to be the block of b consecutive observations starting from Y_i , $i = 1, \dots, n - b + 1$. The moving blocks bootstrap is based on sampling with replacement from the collection $\{\xi_1, \dots, \xi_{n-b+1}\}$. Suppose that k is an integer such that kb is approximately n , and let ξ_1^*, \dots, ξ_k^* be sampled independently and with replacement from $\{\xi_1, \dots, \xi_{n-b+1}\}$. Let the $l = kb$ elements of ξ_1^*, \dots, ξ_k^* be concatenated into a single vector $(Z_1, \dots, Z_l) \equiv (\xi_1^*, \dots, \xi_k^*)$. Then (Z_1, \dots, Z_l) is the bootstrap sample under the moving blocks bootstrap scheme and, for example, a bootstrap estimate of $P\{\sqrt{n}(\bar{Y}_n - \mu) \leq z\}$ is $P\{\sqrt{l}(\bar{Z}_l - \bar{Y}_n) \leq z\}$, where the probability is computed under the moving blocks resampling scheme, and where $\bar{Z}_l = l^{-1} \sum_{i=1}^l Z_i$. Consistency of the distribution estimator under the model of m -dependence is achieved if b is allowed to grow to infinity with n . The classical bootstrap, used for an IID data sample, which independently resamples from the given data points, can be thought of as a special case $b = 1$ of the moving blocks bootstrap.

Question 1 Let Y_1, \dots, Y_n be observations from a univariate m -dependent stationary sequence, and the statistic of interest be the sample mean $\bar{Y}_n = n^{-1} \sum_{i=1}^n Y_i$. The moving blocks bootstrap estimate of the variance of $\sqrt{n}\bar{Y}_n$ is the variance of $\sqrt{l}\bar{Z}_l$ under the moving blocks resampling scheme.

Show that the variance estimator is

$$\frac{b}{n-b+1} \sum_{i=1}^{n-b+1} \left(\frac{1}{b} \sum_{j=i}^{i-1+b} Y_j - E^* \bar{Z}_l \right)^2,$$

where

$$E^* \bar{Z}_l = \frac{1}{n-b+1} \sum_{i=1}^{n-b+1} \frac{1}{b} \sum_{j=i}^{i-1+b} Y_j.$$

Question 2 Simulate Y_1, \dots, Y_{225} from the moving average model

$$Y_t = W_t + W_{t-1} + W_{t-2} + W_{t-3},$$

where the W_t are IID $N(0, 1)$ (You may use any method to generate W_t).

Show that the moving average model satisfies the conditions of m -dependence and find $\text{var}(\frac{1}{15} \sum_{i=1}^{225} Y_i)$. Compute, as a function of b , the moving blocks bootstrap estimate of the variance of $\frac{1}{15} \sum_{i=1}^{225} Y_i$. What choice of b is best? How bad is the classical (IID sampling) bootstrap?

Question 3 Suppose that the cost associated with getting the variance estimate wrong is the square of the error. Repeat the whole simulation exercise, using different randomising seeds, a reasonable number of times. Plot a graph of the average mean square error against b and hence compare the accuracies of the moving blocks variance estimators constructed using different block sizes.

Question 4 Repeat the above questions for increasing values of n . For each n decide what value of b is least costly. Plot the best block size versus n . What general advice would you offer about choice of block size b ?

References

- [1] Efron, B. Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7** (1979) 1–26.
- [2] Künsch, H.R. The jackknife and the bootstrap for general stationary observations. *Ann. Statist.* **17** (1989) 1217–1241.