A MACHINE LEARNING APPROACH TO CHANGE-POINT DETECTION

SANTOSH KANDEL
DEPARTMENT OF MATHEMATICS AND STATISTICS

Introduction

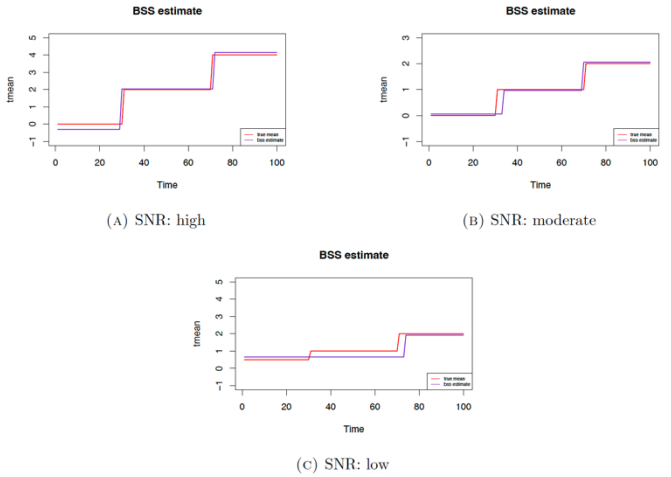
Change-poi nts detection, a main technical tool to analyze a time series data, has applications in many fields such as bioscience, climatology, economics, and engineering. For example, medical doctors monitor physiological time series to identify sudden abnormalities that may relate to a certain medical condition; stock market analysts monitor daily stock market data over a period to see if there are any abnormal shifting; and meteorologists monitor daily average temperatures of a certain location over the years to see if there are any sudden changes. The main objective is to develop a new methodology to study change-point problems inspired by one of the most popular machine learning methods called bagging. This project also assesses this methodology using simulated data. The proposed methodology will be implemented using the programming language R.

Background and Significance

The primary objective of change-point detection is to estimate locations of change-point in the given dataset. It is argued in the literature that identifying changes of a time-evolving statistical quantity may be reduced to identifying changes in the mean of a new data derived from the initial data [2, 3]. In regression analysis, variable selection refers to identifying those independent variables which have the most influence on the dependent variable. From a regression framework standpoint, estimating the number of change points amounts to estimating the number of most influential independent variables. This can be done by using the so-called variable selection method. While variable selection in regression analysis to change-point analysis is an effective method, estimates obtained from this method are expected to be highly unstable [1]. Thus, it is crucial to develop a methodology that addresses this shortcoming of the variable selection method.

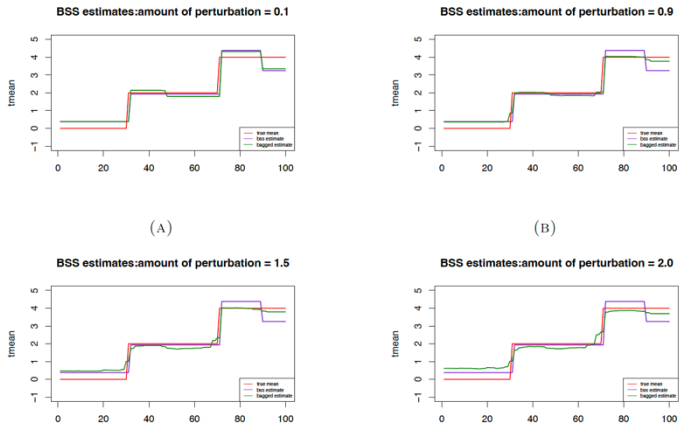
Methodology

Bagging, introduced by Breiman [1], is a method of averaging estimates from a specific statistical procedure. It is performed by generating many data sets from the given dataset by resampling, constructing an estimator from each resampled dataset, and then averaging the estimators. In this study, I propose a different method to generate new data sets and use them for change-point estimates. I will then aggregate the change-points estimates coming from these new data sets as in the bagging.



(A) SNR: high (B) SNR: moderate (C) SNR: low

FIGURE 1. BSS estimates



(A) (B) (C) (D)

FIGURE 2. Bagged BSS estimates

Findings

Our experiment suggests that the performance of variable selection method to change point detection depends on the so-called signal-to-noise ratio (SNR). Some examples of change point estimates are given in Figure-1. The red lines are true means and the purple lines are estimates via variable selection method. In (A) and (B), SNR is relatively high but in (C) SNR is low. We also find that our proposed methodology improves the performance of estimator but again it depends on the SNR (Figure-2, (B), (C) and (D)).

References

[1] L. Breiman, *Bagging predictors*, Machine Learning, 24 (1996), pp. 123–140.
[2] Z. Harchaoui and C. Levy-Leduc, *Multiple change-point estimation with a total variation penalty*, Journal of the American Statistical Association, 105 (2010), pp. 1480–1493.
[3] Y.-C. Yao and S.-T. Au, *Least-squares estimation of a step function*, Sankhya: The Indian Journal of Statistics, Series A, (1989), pp. 370–381.