Data-Driven Tail-Greedy Unbalanced Haar-Fisz Method for Copy Number Alteration Data

Maharani A. Ummi[1], Arief Gusnanto[1], Stuart Barber[1]

[1]Department of Statistics, University of Leeds, Leeds LS2 9JT, United Kingdom

E-mail for correspondence: mmmau@leeds.ac.uk

Abstract: Copy Number Alterations (CNA) are genomic aberrations, in which some regions of a genome exhibit more or less copy number than the normal two. They appear as 'gains' or 'losses' of copy number along the genome and play a key role in cancer diagnosis. In particular, the locations of the gains and losses are of major interest. However, estimation of CNA is a challenging process because CNA data contain inconsistent error variability. Several segmentation methods have been proposed to estimate CNA and many of them perform well for data whose error variability is relatively constant. In practice, real CNA data deviate from this assumption and indicate some dependencies of the variance on the mean value. To address this problem, we have developed a method called Data-Driven Tail-Greedy Unbalanced Haar-Fisz (DDTF) segmentation. The proposed method performs variance stabilization via a Fisz transform to bring the problem into a homoscedastic model before applying a denoising procedure. The use of the Tail-Greedy Unbalanced Haar wavelet also makes it possible to estimate CNA location more precisely compared to the traditional Haar wavelet. Furthermore, our simulation study shows the superiority of DDTF in estimating short segments, which are often difficult to detect by existing methods.

Key words: Copy number alteration; unbalanced Haar wavelet; data-driven denoising; heteroscedasticity; change-point detection

Fryzlewicz P (2008). Data-driven wavelet-Fisz methodology for nonparametric function estimation. Electronic Journal of Statistics 2:863–896.

Fryzlewicz P (2018). Tail-greedy bottom-up data decompositions and fast multiple change-point detection. Annals of Statistics 46.