Change Point Analysis

Given a series of data, change point analysis involves detecting the number and location of change points, locations in the data where some feature, for example the mean, changes. This has applications in a wide variety of areas, including: finance, quality control, genetics, environmental studies and medicine. An example of performing a change point analysis on a randomly generated time series is shown in Figure 1

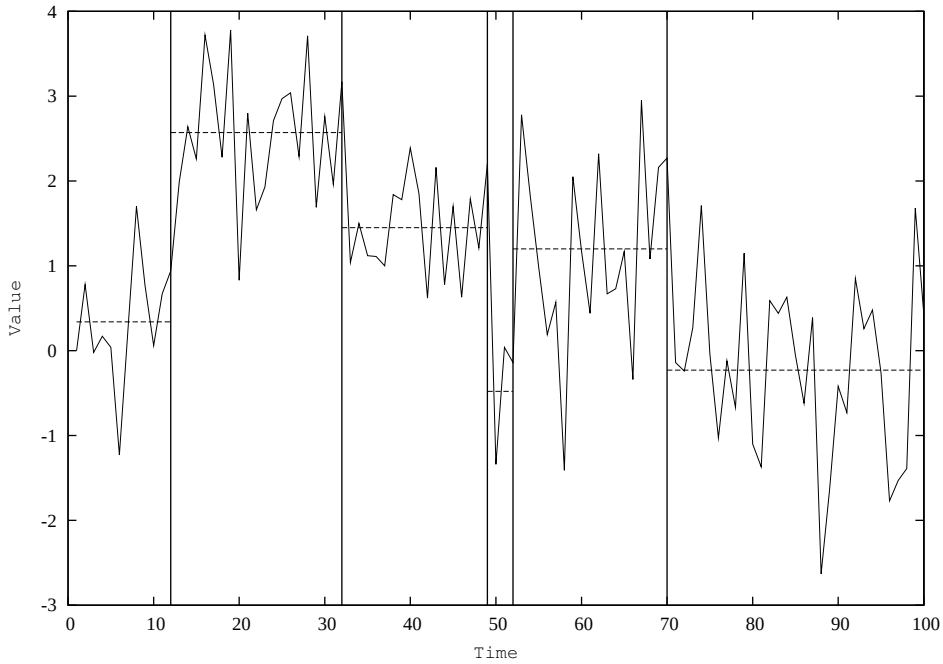

Figure 1: Plot of the estimated location of changes in the mean of a simulated times series, as detected using the PELT algorithm. The vertical lines indicate the location of the change points and the dotted horizontal lines are the estimated mean in each segment.

More formally, let $y_{1:n} = \{y_j : j = 1, 2, \ldots, n\}$ denote a series of data and $\tau = \{\tau_i : i = 1, 2, \ldots, m\}$ denote a set of m ordered change points, with $1 \leq \tau_i \leq n, \tau_m = n$ and $\tau_i \leq \tau_j$ if and only if $i \leq j$ and the equality holds only if $i = j$. For ease of notation we also define $\tau_0 = 0$. Change point detection involves solving:

$$\underset{m,\tau}{\text{minimize}} \sum_{i=1}^{m} \left(C\left(y_{\tau_{i-1}+1:\tau_i}\right) + \beta \right) \tag{1}$$

where $C\left(y_{\tau_{i-1}+1:\tau_i}\right)$ is a cost function and β is a penalty term used to help control the number of change points detected. The m change points, τ, therefore split the data into m segments, with the ith segment being of length n_i and containing $y_{\tau_{i-1}+1:\tau_i}$.

At Mark 25 two methods for detecting change points were introduced into Chapter G13 of the NAG Library; PELT and Binary segmentation. The PELT algorithm of Kilick et al. [2]

is guaranteed to return the optimal solution to equation (1) as long as there exists a constant K such that

$$C\big(y_{(u+1):v}\big) + C\big(y_{(v+1):w}\big) + K \leq C\big(y_{(u+1):w}\big)$$

for all $u < v < w$. Binary segmentation will only give an approximate solution to equation (1) but tends to be quicker, especially for large series with only a few change points.

The cost function, C, can either be chosen from a list of six pre-defined functions (three based on the log-likelihood of the Normal distribution and three based on the log-likelihood of the Gamma, Exponential and Poisson distributions) or supplied by the user.

References

[1] J Chen and A K Gupta. *Parametric Statistical Change Point Analysis With Applications to Genetics (Second Edition)*. Birkhuser, 2010.

[2] R Killick, P Fearnhead, and I A Eckely. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107:500:1590–1598, 2012.