Internship Project

For Students in M2 Data Science, AI, Applied Mathematics

Change-Point Detection For Time Series
A Novel Approach For Efficient Computation With Robust Loss

Keywords: Multiple change-point detection. Dynamic programming. Optimization under constraints. Computational statistics. Rcpp packages.

Supervisors: Vincent RUNGE (Assistant Professor), Guillem RIGAILL (Senior Researcher)
For applying send your CV to vincent.runge@univ-evry.fr

Location: IBGBI buiding, Evry University (member of Paris-Saclay cluster)

When: From May to September (about 5 months). Starting and ending times can be discussed.

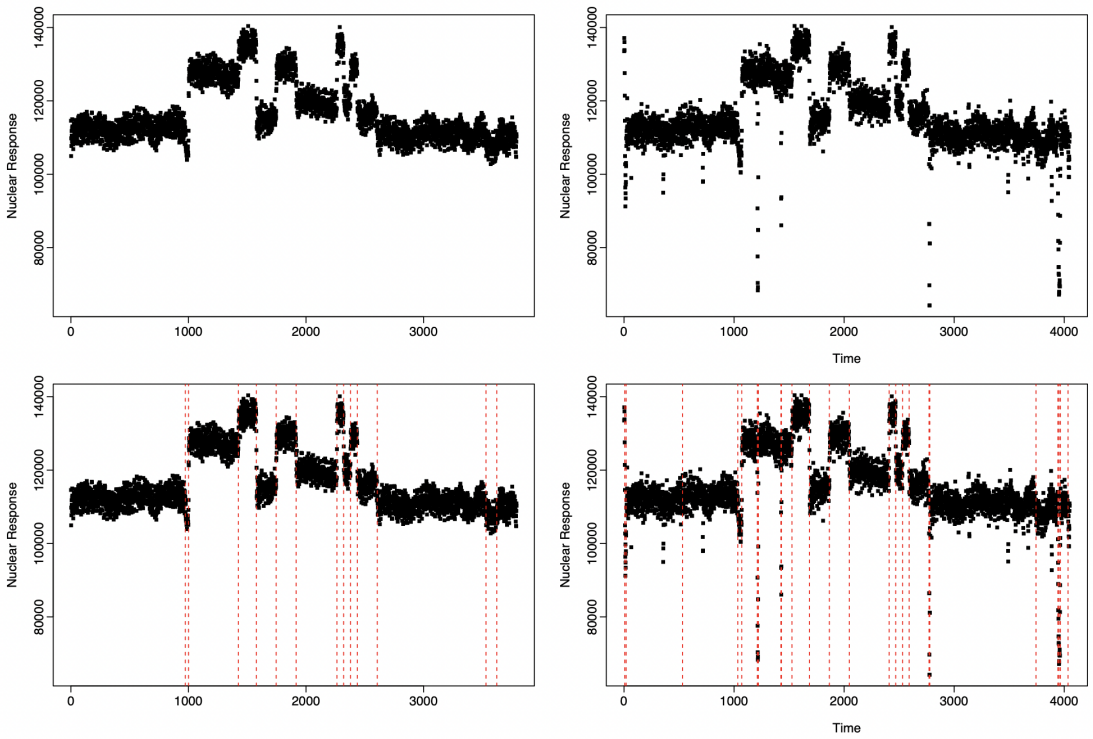

Fig. 1: Two Time series (top) and their segmentation by red vertical lines (bottom). Having outliers in data (right plots) increases the number of changes. A method robust to outliers would be able to ignore these spurious data. Figure taken from [FR19]

In recent years, many methods have been proposed for detecting one or multiple changepoints offline or online in data streams. The reason for such a keen interest in changepoint detection methods lies in its importance for various real-world applications, including bioinformatics, climate and oceanography, econometrics, or finance.

In this project, we consider the problem of detecting multiple change-points in time series contaminated by spurious data (see right side of Figure 1). Dealing with outliers is a tedious task for most machine learning algorithms and often requires the use of some pre-processing steps. However, this poses a risk of losing important information at the initial stage. One solution is to directly employ a robust loss function in the changepoint detection algorithm. Such a loss function can be integrated into a dynamic programming algorithm called FPOP. [Mai+17].

Our goal will be the implementation (R, C++), test and theoretical understanding of a new method capable of dealing with these outliers. The method should be able to discover only the relevant changes while being still time efficient.

A first method, efficient for univariate Gaussian data, has been developed recently [FR19] in our lab. It is based on the update at each time step of a piecewise quadratic function. See Figure 2.

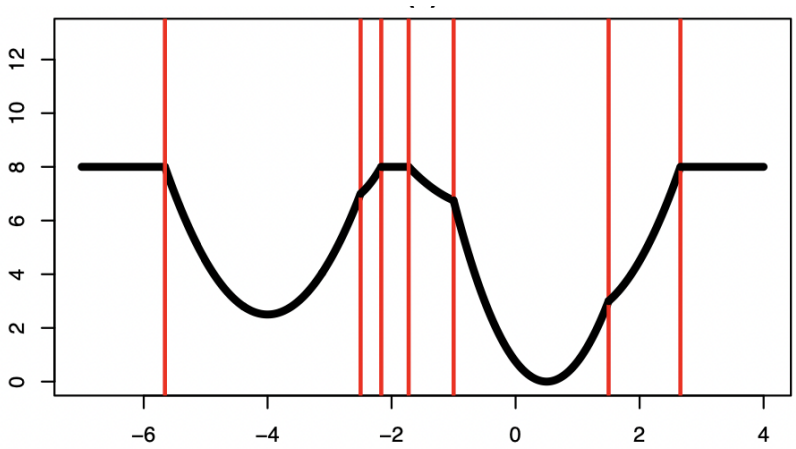

Fig. 2: The functional problem [FR19]. We need to update piecewise quadratic functions at each time step. Data points enter the algorithm one by one, for each new data, the piecewise function is updated.

Our new approach is also based on the functional pruning optimal partitioning (FPOP) algorithm [Mai+17] but uses the solutions of optimization problems under constraints instead of the difficult functional update of Figure 2. With this strategy, more complex models can be considered with a robust loss, e.g. multivariate Gaussian and multivariate Poisson distributions. Importantly, the obtained algorithm is relatively easy to write down.

FPOP dynamic programming algorithm extends the well-known OP and PELT algorithms: see [KFE12] for a good introduction of theses methods.

The intern will be in charge of:

1. Coding the new algorithm in R and C++ (about 50 lines of code only);
2. Leading a simulation study to evaluate its performances (time efficiency, etc);
3. Participating to the theoretical understanding of the algorithm;
4. Writing with its supervisors a short research paper (depending on the obtained results).

References

[KFE12] Rebecca Killick, Paul Fearnhead, and Idris A Eckley. "Optimal detection of changepoints with a linear computational cost". In: *Journal of the American Statistical Association* 107.500 (2012), pp. 1590–1598.

[Mai+17] Robert Maidstone et al. "On optimal multiple changepoint algorithms for large data". In: *Statistics and computing* 27.2 (2017), pp. 519–533.

[FR19] Paul Fearnhead and Guillem Rigaill. "Changepoint detection in the presence of outliers". In: *Journal of the American Statistical Association* 114.525 (2019), pp. 169–183.