Internship – Master 2
Computational statistics – Research project

A changepoint detection algorithm in dimension 2 using pruned dynamic programming

Abstract

We are looking for a motivated M2 student in mathematics or applied mathematics to work on the C++ implementation of an efficient pruned dynamic algorithm for multiple change-point detections.

How long	5–7 months starting in January–March 2018
Lab	LaMME www.math-evry.cnrs.fr/doku.php (45 min from Paris)

Detailed subject

The detection of changepoints is a common problem in data science and time series analysis. The goal is to split the data in contiguous and homogenous segments along time. According to the UK National Research Council (2013), it is at present one of the most challenging and relevant problem for big data applications [1].

We are developping exact dynamic programming approaches for the inference of parametric changepoint models. These algorithms are exact and guaranted to provide the segmentation maximizing the likelihood but they are slower than the often used binary segmentation heuristic. Recently, a new strategy, called functional pruning, has been proposed to speed up calculations [2]. It is currently implemented for models with one parameter (1d) per segment. This internschip is dedicated to the implementation of the functional pruning strategy in dimension 2, allowing the joint/simultaneous segmentation of two time series.

The algorithm to implement during the internship consists in updating a continuous piecewise quadratic function in dimension 2 and updating sets in $\mathbb{R}^2$ obtained by intersecting a number of circles. These sets are illustrated in Figure 1 obtained with the R package plotFPOP. The package can be downloaded on Github using the R code :

```r
devtools::install_github("vrunge/plotFPOP")
library(plotFPOP)
```

The goal of the intership is to implement the algorithm in C++. But if time allows, writing a research paper about the algorithm, new statistical developpements and applications to real genomic data could be considered.

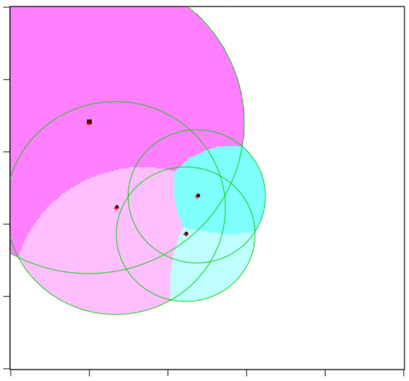

Fig. 1: Trucature of the functional cost in dimension 2 (a piecewise quadratic function)

Skills

- Good knowledge of R, C++ and notions of object-oriented programming
- Not afraid to code, debug and test
- Reading and writing English
- Passionate about statistics
- Interest for biology and genomics in particular
- Last but not least : team spirit and cheerfulness !

Application

Interested individuals should include a half page cover letter describing their experience along with a CV and the names and contact information of one or two references.

`Laboratory` : The internship will be in Évry at the institute IBGBI in the laboratory LaMME.

`Supervisors` : Vincent Runge (Post-Doc), Guillem Rigaill (Researcher - chargé de recherche), Michel Koskas (Researcher - chargé de recherche)

`Duration` : 5 to 8 months between January and August 2018.

`Salary` : about 500 € per month.

`Contact` : Vincent Runge : vincent.runge@univ-evry.fr , Guillem Rigaill : guillem.rigaill@inra.fr

References

[1] National Research Council (2013) Frontiers in massive data analysis *https://www.stat.berkeley.edu/∼ mmahoney/pubs/nrc-massive-data.pdf*

[2] Maidstone, R., Hocking, T., Rigaill, G. and Fearnhead, P. (2017) On Optimal Multiple Changepoint Algorithms for Large Data. *Statistics and computing,* 27:519-533.