Efficient methods for continuous time event recording segmentation

Context

In many scientific areas, more and more data are collected in continuous time: brain activity in neuroscience, camera traps or animal movement recordings in ecology. In many situations, the recorded signal displays abrupt changes, associated with changes in the underlying behavior of the phenomenon under study (volcanic, brain or animal activity). Furthermore, it is often natural to assume the intervals between changes are associated with a limited number of recurrent behaviors (rest or active for neurons, hunting, transit or rest for animal movements). Because such recordings are more and more abundant with larger and larger dimensions, there is a need for automatic, computationally efficient and mathematically grounded methods for systematic data analysis.

Framework

We focus here on the occurrences of one-off events over time, such as volcano eruptions, neuron spikes or animal passages, to name a few. Point processes are a natural framework to describe such datasets. The behavior of a point process is described by its intensity function, which will be supposed to be affected by abrupt changes. Assuming that the observed recording is a realization of a point process over a given period of time, (offline) change-point detection, also called segmentation, precisely aims at locating the times when such changes occurred.

Segmentation has been widely studied for discrete-time data, but the literature in continuous-time is much scarcer. The development of a segmentation method classically requires a (*i*) relevant modeling for the data at hand, (*ii*) a (computationally and statistically) efficient algorithm to locate the change-points and (*iii*) a model selection strategy to choose the number of change-points.
Task (*ii*) is usually achieved by minimizing a contrast between the observed data and the model predictions. This step raises specific optimization difficulties because most contrasts are not convex, nor even continuous with respect to the change-point locations. Designing algorithms capable of determining the optimal set of change-point is therefore a challenging task. In the case of the heterogeneous Poisson process, [Dion-Blanc & al., 2024] recently showed that tasks (*ii*) and (*iii*) can be achieved in a quadratic time, using specific properties of Poisson processes.

Objective

A first aim of the thesis will be to combine segmentation with (unsupervised) classification, that is, to recover both the change-points and the recurrent underlying behavior at once. A possible lead will be to follow [Picard & al., 2007], who proposed a modeling unifying the two purposes in a discrete time context. In this setting, task (*iii*) becomes more complex, as both the number of change-points and the number of recurrent underlying states need to be determined.

It will then be natural to extend the developed methodology to the multivariate case, to deal with simultaneous recordings, such as spikes from several neurons, passages of animals in different neighbor locations, or movements of different animals. The distribution of a multivariate point process then involves interactions between the different recordings. Segmentation then aims at detecting changes in the interaction rules among the different processes.
Many natural phenomena also display self-excitation (or self-inhibition): replicates of earthquakes, or series of animal cries. This induces a time-dependency in the occurrences of the recorded events. The Hawkes process [Hawkes, 1971] provides a generic framework to model such behaviors, but task (*ii*) is then made notoriously more complex, as the algorithmic tools available for (computationally) efficient segmentation do not accommodate such a dependency structure.

Relation to ISCD

The proposed subject is motivated by applications in life and earth sciences and tackles both modeling and algorithmic challenges. The aim is to provide practitioners with mathematically grounded and computationally efficient methods. All the methods to be developed in this thesis will be associated with publications in international journals and with publicly available packages in python or R.

Supervision

The PhD student will be co-supervised by

- Anna Bonnet (https://anna.biogeek.land/),
- Stéphane Robin (HdR, https://scj-robin.github.io/),

from the Laboratoire de Probabilités, Statistique et Modélisation (LPSM), at Sorbonne Université, and

- Emilie Lebarbier (HdR, https://www.parisnanterre.fr/mme-emilie-lebarbier)

from the laboratory MODAL'X at Université Paris-Nanterre, MODAL'X).
The PhD student will be hosted at the LPSM, in Sorbonne Université (Jussieu).

Background

The applicant should have a strong mathematical background, especially in statistics, and good programming skills. A strong interest in applications to life sciences will be more than welcome.

References

- Dion-Blanc, C., Lebarbier, E., & Robin, S. (2023). Multiple change-point detection for Poisson processes. *arXiv preprint arXiv:2302.09103*.
- Picard, F., Robin, S., Lebarbier, E., & Daudin, J. J. (2007). A segmentation/clustering model for the analysis of array CGH data. *Biometrics*, *63*(3), 758-766.
- Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes, Biometrika 58 (1), 83-90