# PhD Proposal

**Titre en français :** Détection de ruptures en lignes pour les séries temporelles multivariées structurées

**Titre en anglais :** Online change-point detection for structured multivariate time series

**Laboratoire de recherche :** Centre Borelli (UMR 9010, CNRS, ENS Paris Saclay, Université de Paris, SSA, INSERM)

**Directeurs de thèse :**

- Laurent Oudre (Professeur des Universités)
- Nicolas Vayatis (Professeur des Universités)
- Argyris Kalogeratos (Senior Researcher)
- Hugo Henneuse (Postdoctoral Researcher)

**Etablissement d'inscription :** ENS Paris-Saclay

**Ecole Doctorale d'inscription :** École Doctorale Mathématiques Hadamard (EDMH)

**Partenaires :** Cette thèse est financée et sera réalisée en collaboration avec les partenaires de l'ANR SCOPED (Université Côte d'Azur, Université de Lorraine, Université de Bordeaux).

**Durée et mode de financement :** 36 mois (100 %)

**Profil recherché :** étudiant(e) titulaire d'un M2 "Recherche" en mathématiques, statistiques ou mathématiques appliquées. Excellent niveau en statistiques, compétences en programmation Python. Expertise en séries temporelles, géométrie et graphes serait un plus.

**Thèmes :** détection de ruptures, séries temporelles, statistiques, espaces non-euclidiens, géométrie riemannienne, groupe de Lie, graphes.

## Context

Change-point detection (CPD) is a fundamental problem in statistics and machine learning, focused on identifying abrupt shifts in the properties of data over time. These shifts, known as change-points, indicate transitions in the underlying distribution or dynamics of a system, which may result from external events or internal structural changes. The objective of CPD is to pinpoint when these changes occur and, in some cases, to understand the nature of the shifts. CPD has a wide range of applications across domains that require online insights and adaptive decision-making, such as medical monitoring, realtime trading, and network security. A growing number of these applications are generating structured, high-dimensional data with non-trivial and intricate geometric properties. These data often display complex relationships and dependencies that go beyond Euclidean spaces, necessitating sophisticated techniques for analysis and interpretation. Prominent examples include time sequences on groups, on manifolds, time sequences of graphs, and graph signals, all of which are central to this project.

A major challenge with dynamic structured data is finding representations that can effectively handle their underlying geometry, which is often defined by application-specific pseudo-distances. A common approach is to embed such data into conventional geometric spaces, like Euclidean spaces, even when they may be more naturally represented in non-Euclidean domains. This mismatch in representation complicates both learning and inference processes. Another challenge is that dynamic-structured data are generated by a variety of sensors and infrastructures that continuously produce, disseminate, and store information. However, this data deluge already surpasses our capacity for analysis and decision-making, necessitating online actions rather than offline processing due to its time-sensitive nature.

**First research lead : online CPD on Lie groups and Riemannian manifolds**

A natural and interesting case study involves Lie groups and Riemannian manifolds, which arise in a wide range of applications. For instance, movement data can be represented as elements of the Lie group of rigid motions, directional data lies on spheres, and angular data on tori. Efficiently detecting changes in such data is therefore an important problem. This task is often delicate, as basic probabilistic notions, such as the mean, may not exist or may not be unique. Even when they can be defined, computing them can be computationally expensive, making standard methods developed for Euclidean data impractical. In this context, identifying functionals that respect the underlying geometry while retaining favorable computational properties becomes a critical challenge.

In this direction, several recent works have proposed methods based on detecting Fréchet mean shifts. Fréchet mean is a generalization of the notion of the mean to non-Euclidean spaces. When data lie on Lie groups or Riemannian manifolds, explicit conditions have been established to ensure the existence and uniqueness of the Fréchet mean, involving the curvature of the underlying space. Furthermore, several algorithms have been proposed to efficiently compute the Fréchet mean on such spaces. These favorable theoretical and computational properties make Fréchet mean shifts a promising alternative to approaches based on Euclidean embedding. However, existing methods are either purely numerical or have primarily focused on the offline setting, motivating the development and statistical analysis of their online counterparts.

Yet, moving from an offline to an online perspective entails significant changes, particularly in terms of statistical analysis. In the offline setting, the goal is to accurately estimate the number of changes that have occurred over a given period and to locate them. This is often formalized by establishing the consistency of the estimators for both the number and locations of the changes, as well as studying their associated convergence rates. In contrast, the online setting focuses on detecting a change as quickly as possible using only past and current observations, while simultaneously controlling the risk of false detections. This is typically formalized by balancing two quantities: the detection delay, which measures the time lag between the actual change and its detection, and the average run length to false alarms, which is the expected number of observations before a false alarm occurs. These two objectives inherently induce a trade-off. This change of paradigm thus implies a natural shift from an estimation perspective to a more test-based framework.

Beyond methods based on Fréchet means, a promising exploratory direction for this thesis is to identify other relevant geometric descriptors in the context of change point detection. For example, one could consider alternative centrality measures adapted to non-Euclidean spaces, such as the Tukey median. Tools from Topological Data Analysis, particularly persistent homology, have also recently attracted significant interest within the time series analysis community and could provide appealing alternatives. In the more specific context of Lie-group-valued time series, several recent works have highlighted the advantages of methods based on path signatures, among other possibilities.

**Second research lead : online CPD on undirected graphs and data over graphs**

As already stated, online CPD tries to detect a change point (CP) the soonest possible, neither having the luxury of storing arbitrary a large amount of information from the stream, nor responding late. In this context, the detection problem is examined at each instant of time t, where there is generally enough data prior to t, while a hypothetical change will only start being manifested after t. Timely detection requires addressing the statistical challenges raised by the scarceness and imbalance of the observations pre- and post-CP. For instance, statistical testing offers typical several tools in the CPD toolbox, such as the likelihood-ratio (or density-ratio) test, however those tools struggle to handle properly such scenarios and lead to non-negligible or even substantial detection lags.

Considering that, at the early post-CP moments, the new data distribution has not formed yet sufficient density mass compared to the pre-CP data distribution, one naturally thinks that CPD can be seen as an anomaly detection task. For that, density-based or model-based isolation methods can be employed, whereas providing decisions with well-characterized statistical properties may require to formalize such approaches within a more thorough framework (e.g. as statistical hypothesis testing). From another angle, recent advances in statistical machine learning have produced methods for non-parametric online likelihood-ratio estimation using advanced kernel methods (such as vector-valued kernels), which can be further developed in regard to their computational, data-requiring features, precision features. For instance, kernel-based methods rely on the build-up of a dictionary of pre-CP observations. To that end, new observations are incorporated in the dictionary in an online fashion as long as they are sufficiently different to the existing elements. This type of strategy aims at keeping the redundancy of the dictionary low, which would maximize the representation power of the induced Reproducible Kernel Hilbert Space while keeping the computational complexity low. Improving on this kind of online dictionary learning can also be highly valuable for a subsequent CPD task.

The previous elements are generic, hence they could handle both univariate and multivariate time-series or data streams. This project, though, aims at incorporating structural information from the data, which is mostly comprehensible if thought as a form of a connectivity graph. The graph can be spatial, meaning that it may correspond to the physical arrangement of a system (e.g. geographical locations of sensors), or a statistical derivative (e.g. a covariance matrix). Whichever the case, the graph defines a discrete topological space in which data are observed, and which is important to be incorporated in the inference or detection process. One frontier that needs to be challenged is how to define suitable isolation mechanisms for data over graphs (e.g. univariate graph signals, or multivariate graph data) that could quickly identify the beginning of a change by taking into account the spatiotemporal effects in the data. Another important frontier is to combine online with graph-based likelihood-ratio estimation, that in turn can produce efficient statistical machine learning methods for the CPD task. Although advances have been made in those last two sides, yet their combination is not trivial at all; further, the adaptation to the CPD problem tops up additional challenges.

Finally, a direction of work will attempt to deal with the CPD task on graph streams. In that case, the “signal” is the stream of graph structures. This problem has been addressed by statistical testing methods. The offline and even more the online versions of this problem are both into the perimeter of the objectives of this project. Moreover, there are very interesting lines to be drawn between the graph stream case and the case where a signal (univariate or not) is seen as a dynamic graph.