

## Internship

### Model-Based Multivariate Time Series Analysis

**Level:** Masters or Engineer level (specialized in applied mathematics / statistics / computer science)

**Supervisors:** Etienne Goffinet ([etienne.goffinet@lipn.univ-paris13.fr](mailto:etienne.goffinet@lipn.univ-paris13.fr)), Hanane Azzag ([hanane.azzag@lipn.univ-paris13.fr](mailto:hanane.azzag@lipn.univ-paris13.fr)), Mustapha Lebbah ([mustapha.lebbah@lipn.univ-paris13.fr](mailto:mustapha.lebbah@lipn.univ-paris13.fr))

**Duration:** 5 or 6 months, starting March/April 2021 (flexible)

**Location:** Université Sorbonne Paris Nord, LIPN - UMR 7030 – CNRS

#### Context

Most real-world data have a temporal component, whether it is measurements of natural processes (weather, sound waves) or man-made (sensors, medical data, stock market). Analysis of time series data has been the subject of active research for decades and is considered as one of the top 10 challenging problems in data mining due to its unique properties.

Unsupervised machine learning on time series has gained interest recently, but is still a research field, especially concerning multivariate time series. In this domain, several approaches have been recently developed based on mixture models. In particular, Functional Latent Block Model which is a model-based framework that simultaneously discriminates interesting variables and creates meaningful clusters of observations.

This internship is an opportunity to study the domain of model-based multivariate time series analysis, and develop dedicated tools in collaboration with R&D teams of Groupe Renault. The skills acquired during this internship will be extremely valuable for working both in research and industry.

#### Subject

The internship will consist in 3 phases:

1. Study the current state-of-the-art on time series analysis, with a focus on model-based unsupervised learning, clustering and visualization.
2. Review the current state-of-the-art software tools and architectures for large-scale time series analysis. This includes useful software packages and libraries to build, train and deploy models on time series data; data visualization tools; large-scale storage and databases.

3. Based on previous studies, implement one or several algorithms or tools that do not exist yet. Preferred language used is Scala, but Python can also be used. The results obtained during the internship may lead to contributions to open-source, or even a scientific publication, depending on the intern's skills and motivation.

## Required skills

Solid mathematical background (Bs/Ms), applied mathematics, statistics, signal processing...

CS background: algorithms, complexity theory. Good imperative and object-oriented programming skills. Functional programming would be a plus.

Proficiency in at least one multi-purpose language, including (but not limited to): Scala (preferred), Java, Python, C++, Go...

Knowledge of one or more of the following technologies: Apache Spark (preferred); SQL/NoSQL databases; version control (git).

Scientific computing and ML libraries: Python (numpy, pandas, scikit-learn...), Scala (breeze, smile...).

In the current health context, the candidate is expected to be autonomous, organized and able to adapt with remote work methodology.

## Bibliography

- Govaert, G., & Nadif, M. (2013). *Co-clustering: models, algorithms and applications*. John Wiley & Sons.
- Melnykov, V. (2016). Model-based biclustering of clickstream data. *Computational Statistics & Data Analysis*, 93, 31-45.
- Chamroukhi, F., & Biernacki, C. (2017, July). Model-Based Co-Clustering of Multivariate Functional Data.
- Slimen, Y. B., Allio, S., & Jacques, J. (2018). Model-based co-clustering for functional data. *Neurocomputing*, 291, 97-108.

This is a paid internship: 577.50€/month.

Contact and other information to apply, please send a resume, cover letter and transcripts of your last two years to [etienne.goffinet@lipn.univ-paris13.fr](mailto:etienne.goffinet@lipn.univ-paris13.fr).