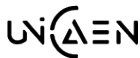




Projet RIN AStERiCs

asterics.lmno.cnrs.fr

FAICEL CHAMROUKHI



Journée du Pôle **Sciences du Numérique**

Caen, 2 octobre 2018



AStERiCs (depuis avril 2018, pour 2 ans)

Apprentissage Statistique à l'Echelle pour la Représentation et la Classification non-supervisées

Partenaires : Région Normandie; Univ Caen - LMNO; Univ Rouen - LMRS

Université de Caen, UMR LMNO

- Faïcel Chamroukhi
- Jalal Fadili
- Christophe Chesneau
- André Sesboüé



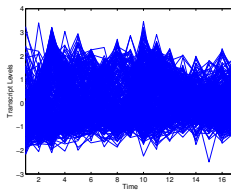
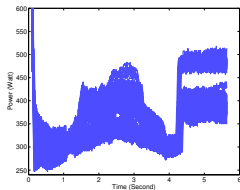
Université de Rouen, UMR LMRS

- Antoine Channarond
- Gaëlle Chagny
- Nicolas Vergne
- Caroline Bérard

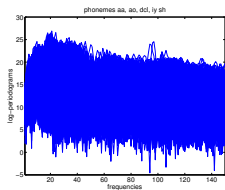


↔ Relève de l'**Axe Sciences des Données** du pôle SN

Données longitudinales de plus en plus fréquentes

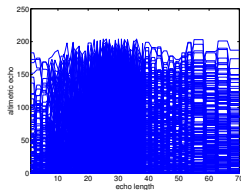


Railway switch curves



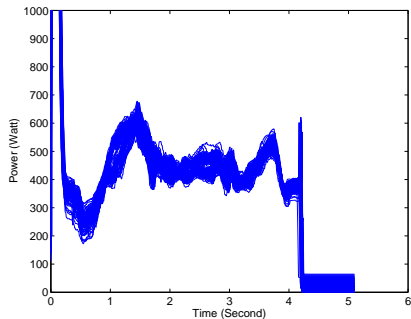
Phonemes curves

Yeast cell cycle curves

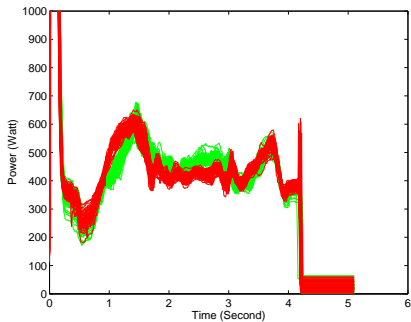


Satellite waveforms

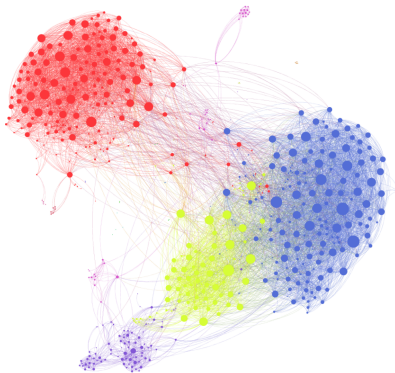
Clustering/segmentation de données temporelles



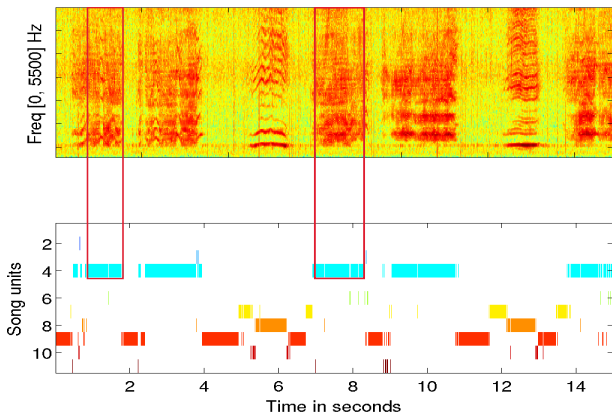
Clustering/segmentation de données temporelles



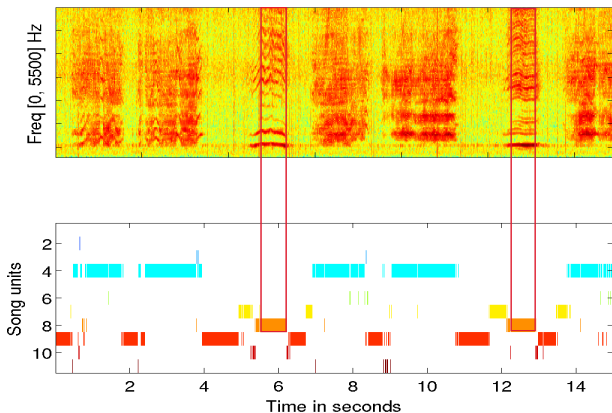
Clustering de données représentées par des graphes



Décomposition parcimonieuse non-supervisée



Décomposition parcimonieuse non-supervisée



Verrous scientifiques

- Données de grande dimension : recours à des modèles parcimonieux ; de données longitudinales, de séries temporelles ;
- Hétérogénéité des informations à extraire, sans ou avec très peu d'annotations : nécessité de modèles non-supervisés ;
- Organisation multi-échelles des données (groupes, hiérarchie de groupes, etc) : modélisation hiérarchique ;
- Représentation, annotation et manipulation de données hétérogènes en fonction de points de vue différents (chercheurs, entreprises, etc) : généralité des modèles ;
- Utilisation effective des traitements par les différents acteurs : nécessité d'algorithmes d'analyse rapides

Objectifs

- Données complexes \leftrightarrow *hétérogènes, temporelles dynamiques, fonctionnelles, incomplètes, de grande dimension, et disponibles en masse*
- **Objectif** : Transformation de telles données en connaissances :
 \leftrightarrow Reconstruction/révélation de structures cachées, i.e, (hiérarchie de groupes ; sélection de variables et prédiction, etc

\leftrightarrow AStERiCs vise à élaborer un cadre scientifique et technique pour traiter et analyser des données massives hétérogènes et peu ou non-annotées

\leftrightarrow Avec une visibilité à l'international

Axes du projet

- 1 Modélisation non supervisée par des modèles à variables latentes (MVL)
- 2 Inférence efficace non supervisée à grande échelle des MVL
- 3 Prototypage des algorithmes développés

Modélisation statistique par des MVL à l'échelle

Cadre scientifique général

↪ **Modèles** statistiques à **structure latente** : $f(x|\theta) = \int_{\mathcal{Z}} f(x, z|\theta) dz$

↪ **Inférence** à grande échelle par régularisation et échantillonnage :

$$\hat{\theta} \in \arg \max_{\theta} \ell(\theta) - \text{Pen}_{\lambda}(\theta)$$

Modélisation statistique non supervisée à grande échelle par des MVL

- Apprentissage génératif, via des modèles à variables latentes (**régression** et **classification**).
- représenter explicitement la structure des données brutes et la révéler
 - ↪ \exists fondement théorique solide
 - ↪ Outils afférents d'estimation et de choix de modèle
- \Rightarrow n'ont pas été considérés avec succès pour une analyse à grande échelle

Inférence non supervisée à grande échelle des MVL

Inférence en grande dimension

- L'inférence se ramène en général à l'optimisation de problèmes non linéaires complexes. à grande échelle :
 - ↪ suggère de nouvelles stratégies de **régularisation** pour pallier la grande dimension
 - ↪ Méthodes **parcimonieuses** pour une meilleure représentation

Données de gros volume

- la distribution des calculs est une façon naturelle de s'y prendre
- méthode : échantillonnage et inférence des modèles agrégés à partir d'un gros volume de données
- ↪ Nouvelles stratégies d'agrégation des estimateurs et de sélection de modèle

Développement d'une plateforme logicielle BigData

- plateforme scientifique et technique en libre accès
- mise à disposition d'algorithmes propres pour de données hétérogènes de différents types (temporelles, graphes, bioacoustiques, etc).
- permettrait également de participer à faire évoluer l'offre Master autour du thème du projet (projets, challenges etc)

Actions prévues

- Intégration d'algorithmes propres déjà développés et ce sur de diverses applications réelles
- Intégration des algorithmes qui seront développés dans le cadre projet

Ressources de calcul HPC



Membres

Permanents

- Faïcel Chamroukhi (Université de Caen, UMR LMNO)
- Jalal Fadili
- Christophe Chesneau
- André Sesboüé
- Antoine Channarond (Université de Rouen, UMR LMRS)
- Gaëlle Chagny
- Nicolas Vergne
- Caroline Bérard

Recrutements : Unicaen - LMNO

- Postdoc José Gomez (Thèse Cergy) : avril/2018 - pour 18 mois
- IGR Marius Bartcus (Thèse Toulon) : septembre/2018 - pour 18 mois

Recrutements : Univ Rouen - LMNO

- Postdoc VanHà Hoang (Thèse Lille) : sept/2018 - pour 18 mois
- IGR à venir (recherche en cours) - pour 14 mois

Calendrier prévisionnel du projet

Table2. Calendrier du projet et dates des recrutements :

Tâche	Partenaire responsable		Année 1				Année 2				
	LMNO	LMRS	T0+3	T0+6	T0+9	T0+12	T0+15	T0+18	T0+21	T0+24	
Tâche 0 (coordination)											
Tâche 1.1			PostDoc 18mois), Recrutement à t0+6								
Tâche 1.2			Postdoc (18mois), Recrutement à t0								
Tâche 2.1			IGR (18mois) , Recrutement à t0								
Tâche 2.2							1 IGR (14mois), Recrutement à t0+10				
Etat d'avancement							Livrable 1				Livrable 2
Etat des dépenses							Etat 1				Etat 2

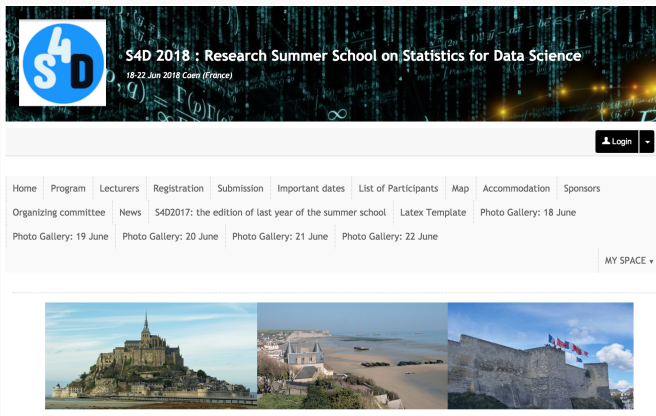
Recherche en cours d'un IGR de 14 mois

Projet complémentaire

- AStERiCs : 2018-2020 est complémentaire avec le projet ANR SMILES (nov/2018 -pour 42 mois)
- Partenaires SMILES : UMR LMNO et UMR LMRS (Normandie), UMR LIS (Paca), INRIA-Modal (HdF)
↔ Dénominateur commun :
Apprentissage de modèles à variables latentes
+ Plateforme logicielle d'AStERiCs

Fait marquant

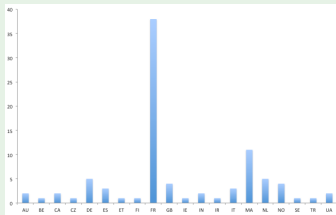
Ecole de recherche internationale à Caen : <http://s4d.sciencesconf.org>



The screenshot shows the header of the S4D 2018 website. On the left is a circular logo with 'S4D' in white on a blue background. To its right, the text reads 'S4D 2018 : Research Summer School on Statistics for Data Science' and '18-22 Jun 2018 Caen (France)'. The background of the header is a dark green field with mathematical formulas and symbols like ∞ , π , and σ . Below the header is a navigation menu with links: Home, Program, Lecturers, Registration, Submission, Important dates, List of Participants, Map, Accommodation, Sponsors, Organizing committee, News, S4D2017: the edition of last year of the summer school, Latex Template, Photo Gallery: 18 June, Photo Gallery: 19 June, Photo Gallery: 20 June, Photo Gallery: 21 June, Photo Gallery: 22 June, and MY SPACE.

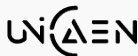
- Talks covering both tutorial and advanced aspects at the interface of Statistics, Machine Learning and Optimization \leftrightarrow the main data science fields
- Theoretical foundations and algorithmic aspects, as well as typical case studies in complex and large-scale scenarios

89 participants (nombre de places limité) de 20 pays



- 2ème édition à Caen. (prochaine édition à l'étranger)
- Budget 18K euro (dont 7K euro sur le RIN)

Sponsors



Publications

- Faicel Chamroukhi and Hien D. Nguyen. Model-based clustering and classification of functional data. 2018. URL <https://chamroukhi.users.lmno.cnrs.fr/papers/MBCC-FDA.pdf>. arXiv:1803.00276v2
- Faicel Chamroukhi and Bao T. Huynh. Regularized mixtures of experts for high-dimensional data. 2018. URL <https://chamroukhi.users.lmno.cnrs.fr/papers/RMoE-HDD.pdf>
- Hien D. Nguyen and Faicel Chamroukhi. Practical and theoretical aspects of mixture-of-experts modeling: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, pages e1246–n/a, Feb 2018. ISSN 1942-4795. doi: 10.1002/widm.1246. URL <http://dx.doi.org/10.1002/widm.1246>
- F. Chamroukhi and B-T. Huynh, "Regularized Mixture of Experts for High-Dimensional Data", Journées de Statistique, Juin 2018, Paris
- F. Chamroukhi and B-T. Huynh, "Regularised maximum-likelihood estimation of Mixture-of-experts" , In proceedings of the IEEE International Joint Conference on Neural Networks, July 2018, Rio.
- M Sautreuil, C Bérard, G Chagny, A Channarond, A Roche, N Vergne. Modèle de mélange binomial négatif bivarié pour l'analyse de données RNA-Seq. Journées de statistique. Juin 2018, Paris

AStERiCs : Apprentissage Statistique à l'Echelle pour la Représentation et la Classification non-supervisées

Projet de recherche Réseaux d'Intérêts Normand.

Contexte et présentation générale du projet:

La disponibilité des données en masse révolutionne les questions relatives à leur traitement, analyse, exploitation et valorisation par les acteurs du numérique (académiques, entreprises, acteurs politiques, etc). La problématique principale est celle de l'élaboration de modèles originaux et génériques permettant la représentation et la classification de données massives, et celle du développement d'algorithmes efficaces optimisés à l'échelle. Ce contexte de traitement et d'analyse à grande échelle rompt en effet avec la façon selon laquelle se posait classiquement la question de la construction et de l'inférence des modèles à partir de données brutes; la plupart de ceux de l'état de l'art se trouvent en effet inopérants à l'échelle, aussi bien d'un point de vue théorique, que pratique : problèmes d'inférence d'un très grand nombre de paramètres (fléau de la dimension), et/ou incapacité en temps et/ou en mémoire de mettre en œuvre des algorithmes centralisés classiques pour de très gros volumes de données, etc.

AStERiCs

Est un projet de recherche fondamentale financé dans le cadre du dispositif RIN (Réseaux d'Intérêts Normands)-Recherche dont l'objectif structural est de fédérer la recherche scientifique en Normandie dans le domaine de la science statistique des données, en s'appuyant sur une démarche scientifique pluridisciplinaire impliquant modélisation mathématique, inférence, représentation et classification de données issues d'environnements complexes, hétérogènes, dynamiques et incertains. AStERiCs vise à élaborer un cadre scientifique et technique, complet, pour traiter, analyser, exploiter et valoriser des données massives, complexes, hétérogènes, dynamiques et peu ou non-annotées. Le but est de transformer des données en connaissances sous forme de représentations précises des informations liées aux données, de catégorisations pertinentes de telles informations, jusqu'à la valorisation de celles-ci en révélant/restaurant le modèle générateur des données. Le projet AStERiCs traite ainsi le problème de la grande échelle, sous tous ses aspects de modélisation et d'inférence. Plus précisément, les axes de

Members

Faïcel Chamroukhi (PR, PI)
Antoine Channaron (MCF)
Jalal Fadili (PR)
André Sesboué (MCF)
Christophe Chesneau (MCF)
Gaëlle Chagny (CR CNRS)
Caroline Bérard (MCF)
Nicolas Vergne (MCF)
Jose Gregorio Gómez García (Postdoc)
Marius Bartcus (Postdoc)
Van Ha Hoang (Postdoc)



RÉGION
NORMANDIE
Financé par la région normandie