Sujet 1 : Clustering spectral

Ce sujet devra présenter l'algorithme de clustering spectral. Il s'agira

- tout d'abord d'expliquer la génèse de cette méthode (basée sur les graphes) à partir du tutoriel de Ulrich von Luxburg disponible ici : `https://arxiv.org/abs/0711.0189`.
- d'illustrer ensuite la méthode sur R ou Python en utilisant des données réelles ou simulées. Sur R, on pourra par exemple utiliser la fonction `specc` du package `kernlab`. On pourra utiliser ces illustrations pour rappeler l'astuce du noyau dans ce cadre.

Sujet 2 : Clustering en grande dimension

Ce sujet devra étudier une famille d'algorithmes de clustering pour la grande dimension présenté dans l'article

- Bouveyron, C. Girard, S. and Schmid, C. (2007) "High-Dimensional Data Clustering", Computational Statistics and Data Analysis, vol. 52 (1), pp. 502–519

On pourra illustrer la méthode à l'aide du package R `HDclassif` présenté dans le papier

- Berge, L. Bouveyron, C. and Girard, S. (2012) "HDclassif: An R Package for Model-Based Clustering and Discriminant Analysis of High-Dimensional Data", Journal of Statistical Software, 46(6), 1–29, url: http://www.jstatsoft.org/v46/i06/

Sujet 3 : Compléments lasso

Dans ce projet, il faudra faire de la recherche bibliographique pour présenter des notions qui ont été très peu abordées dans le cours sur la partie lasso, notamment :

1. algorithme(s) permettant de calculer les estimateurs LASSO, par exemple l'algorithme LARS et/ou descente de coordonnées. On pourra présenter ces algorithmes, les expliquer et éventuellement les programmer.

2. les variantes des pénalités ridges/lasso, par exemple Group Lasso, Fused Lasso, Sparse Group Lasso, ... Il s'agira d'expliquer les cadres propices à l'utilisation de ces pénalités et de les illustrer sur logiciel avec des données réelles ou simulées.

Sujet 4 : Biclustering

Le biclustering est une technique d'apprentissage statistique qui produit simultanément des clusters en se basant sur les lignes (individus) et colonnes (variables) d'un jeu de données. Ce thème est motivé ainsi dans le document de Sébastian Kaiser (que l'on pourra trouver ici : `https://edoc.ub.uni-muenchen.de/13073/1/Kaiser_Sebastian.pdf`) : "A typical situation to calculate bicluster are a high dimensional dataset with many variables, so that normal cluster algorithms lead to diffuse results due to many uncorrelated variables. Also biclustering is useful if there is a assumed connection of objects and some of the variables in the dataset, e.g. some objects have 'similar' patterns for a given set of variables." On trouve de nombreuses applications du biclustering, notamment en génétique mais aussi dans le domaine de la recommandation où l'objectif est souvent d'identifier des groupes d'individus qui ont aimé ou consommé certains produits.

Le projet consistera à présenter quelques algorithmes de clustering, à les motiver sur des problématiques de la vie réelle et à les mettre en œuvre sur logiciel. On pourra par exemple utiliser le package `biclust` de R.

Sujet 5 : Splines et/ou ondelettes (plus technique)

Ce sujet méthodologique reviendra à présenter les méthodes de splines (régression ou lissage) et/ou d'ondelettes. Pour le thème choisi, il conviendra de définir proprement les objets mathématiques et de les illustrer sur logiciel sur des données réelles ou simulées.