Stage de M2 : heuristique de pente pour la sélection de variables en grande dimension.

(English version below)

Encadrant.e.s et contacts :

- Perrine LACROIX, Laboratoire de Mathématiques Jean Leray, Nantes Université, perrine.lacroix@univ-nantes.fr
- Vincent BRAULT, Laboratoire Jean Kuntzmann, Université Grenoble Alpes, vincent.brault@univ-grenoble-alpes.fr

Contexte :

Dans le cadre des modélisations, les statisticien·ne·s doivent parfois choisir entre plusieurs modèles possibles. Ce cas se rencontre par exemple pour les modèles linéaires où la variable Y est expliquée par une sous partie des variables descriptives disponibles $X_1, \cdots, X_p$; en particulier si p est plus grand que n (voir Tibshirani, 1996); un modèle est alors un sous-ensemble de variables parmi les p que l'on doit choisir judicieusement. Nous rencontrons aussi cette question dans le cas des modèles de mélanges où plusieurs lois sont présentes mais nous ignorons le nombre (voir par exemple Dempster et al., 1977) ou dans d'autres modélisations comme les processus *ARMA* en série chronologique (voir par exemple Droesbeke et al., 1989). Pour répondre à la question du choix du modèle, il est possible d'utiliser des critères de sélection de modèles comme l'*Akaike Information Criterion* (AIC) de Akaike (1973) (qui pénalise la vraisemblance par 2 fois le nombre de paramètres) ou le *Bayesian Information Criterion* (BIC) de Schwarz (1978) (qui pénalise la vraisemblance par une fonction du nombre d'observations) ; les résultats étant différents suivant la forme de la pénalité et répondant à des objectifs différents (prédiction pour l'AIC et consistance pour le BIC).
La question de la calibration de cette pénalité est ardue et a mené à plusieurs théories qui ont permis d'introduire des critères plus complexes mais plus judicieux que l'AIC et le BIC. Une de ces théories, étudiée dans ce stage, s'appelle l'heuristique de pente dont le principe est le suivant. Étant données une collection de modèles et une fonction objective appelée *contraste* (par exemple, les moindre carrés pour une estimation linéaire), la théorie montre que si nous conservons le modèle minimisant le contraste empirique (fonction des observations), nous choisirons quasiment à chaque fois le modèle le plus complexe (on parle de sur-apprentissage) ;

l'idée étant qu'à partir d'un moment, nous n'améliorons plus le modèle mais le côté aléatoire des données. Ainsi, il est nécessaire de pénaliser ce contraste par une fonction (appelée pénalité et dépendante de la complexité du modèle) souvent connue à constantes près. La méthode de l'heuristique de pente consiste à calibrer ces constantes inconnues directement sur le jeu de données disponible en exploitant un certain comportement affine (théoriquement prouvé) existant sur les gros modèles de la collection. Cette heuristique est fondée sur les travaux théoriques de (Birgé and Massart, 2007) et est appliquée dans de nombreux modèles statistiques (Baudry et al., 2012).
A ce jour, l'heuristique de pente est codé dans le package **R** *Capushe* (Brault et al., 2012) uniquement lorsque la pénalité est connue à une seule constante inconnue près et nécessite souvent un grand nombre d'observations pour améliorer l'estimation des contrastes. L'idée du stage est de généraliser le code en implémentant une version plus robuste et/ou multivariée. Nous pourrons par exemple nous intéresser à des méthodes de ré-échantillonnages de type bootstrap et nous inspirer des travaux de (Lacroix, 2022) (chapitre 5) pour une première approche. La personne effectuant ce stage pourra alors étudier ces principes suivants deux modélisations différentes : les modèles linéaires sparses et les modèles de mélanges.

Objectifs du stage :

Les objectifs du stage sont les suivants :

- comprendre le principe de sélection de modèle et de l'heuristique de pente,
- s'approprier l'algorithme implémenté dans le package **R** *Capushe*,
- étudier des généralisations de la méthode de l'heuristique de pente : la rendre plus robuste, la rendre multivariée, etc...
- implémenter les nouvelles méthodes,
- évaluer leur qualité de prédiction via une large étude de simulations pour différents modèles : régression linéaire, détection de rupture, modèles de mélange, modèles graphiques, etc...

Mots-clés :

sélection de modèle, statistique en grande dimension, prédiction, sélection de variables, pénalisation, heuristique de pente.

Profil :

Etudiant.e de M2 (ou équivalent) en statistique, en sciences des données ou en dernière année d'école d'ingénieur avec :

- de compétences solides en statistique, en particulier en statistique en grande dimension,
- des connaissances en modélisations linéaires sparses et/ou en modèle de mélange sont un plus mais ne sont pas des pré-requis,
- des compétences en programmation avancée sur **R**.

Détails pratiques :

- Localisation du stage : le stage sera co-encadré par Perrine Lacroix et Vincent Brault, et sera basé au Laboratoire de Mathématiques Jean Leray (Nantes), avec la possibilité d'un séjour court au Laboratoire Jean Kuntzmann (Grenoble).
- Période du stage : 6-mois, printemps et été 2026. Le stage démarrera dès la fin du programme de master de l'étudiant.e. A noter que le LMJL sera fermé administrativement du 24 Juillet au 26 Août 2026 et que l'étudiant.e. pourra prendre des vacances conformément au cadre légal.
- Gratification du/de la stagiaire suivant la grille officielle de la fonction publique : environ 650 euros net/mois [1]. Ce stage est financé par l'académie PULSAR qui est l'académie des jeunes chercheurs des Pays de la Loire.
- Pour postuler : les candidat.e.s intéressé.e.s peuvent envoyer un CV et une lettre de motivation[2] à Vincent Brault et Perrine Lacroix.

[1] Moyenne dépendant du nombre de jours effectivement travaillés dans le mois

[2] Tout dossier généré à l'aide d'un grand modèle de langage risque peut être rejeté immédiatement avant de parvenir aux encadrant·e·s.

M2 internship : slope heuristics for high-dimensional variable selection.

Supervisors with contacts :

- Perrine LACROIX, Laboratoire de Mathématiques Jean Leray, Nantes Université, perrine.lacroix@univ-nantes.fr
- Vincent BRAULT, Laboratoire Jean Kuntzmann, Université Grenoble Alpes, vincent.brault@univ-grenoble-alpes.fr

Context :

In modeling frameworks, statisticians usually have to choose between several possible models. For example, this is the case for linear models where the variable Y is explained by a subset of the set of the available descriptive variables $X_1, \cdots, X_p$; especially when p is greater than n (see Tibshirani, 1996) ; a model is then a subset of variables among the p that have be judiciously chosen. We also encounter this issue in the case of mixture models where several distributions are present but we do not know the number (see for example Dempster et al., 1977) or in other models such as *ARMA* in time series processes (see for instance Droesbeke et al., 1989). To answer the issue of model selection, we can use model selection criteria such as the *Akaike Information Criterion* (AIC) from Akaike (1973) (where likelihood is penalized by 2 times the number of parameters) or the *Bayesian Information Criterion* (BIC) from Schwarz (1978) (where likelihood is penalized by a function of the number of observations) ; the results differ depending on the form of the penalty and respond to different objectives (prediction for AIC and consistency for BIC).
The question of penalty calibration is difficult and has led to several theories that have introduced more complex but more appropriate criteria than AIC and BIC. One of these theories, studied in this internship, is called the slope heuristic, whose principle is as follows. Given a collection of models and an function called contrast (for example, least squares for a linear estimate), the theory shows that if we keep the model that minimizes the empirical contrast (function of observations), we will almost always choose the most complex model (this phenomenon is called overfitting) ; the idea being that at a certain point, we are no longer improving the model but rather the randomness of the data. Thus, it is necessary to penalize this contrast with a function (called a penalty and dependent on the complexity of the model) that is often known up to constants. The slope heuristic method involves calibrating these unknown

constants directly on the available dataset by exploiting a certain affine behavior (theoretically proven) that exists in the large models in the collection. This heuristic is based on the theoretical work of (Birgé and Massart, 2007) and is applied in many statistical models (Baudry et al., 2012).
To date, slope heuristics have been coded in the **R** *Capushe* package (Brault et al., 2012) only when the penalty is known up to a single unknown constant and often requires a large number of observations to improve contrast estimation. The idea of the internship is to generalize the code by implementing a more robust and/or multivariate version. For example, we could test bootstrap-type resampling methods and draw inspiration from the work of (Lacroix, 2022) (chapter 5) for an initial approach. The person doing this internship will then be able to study these principles using two different models : sparse linear models and mixture models.

Internship objectives :

The objectives of the internship are :

- understand the model selection principle and the slope heuristics method,
- use the algorithm implemented in the **R** package *Capushe*,
- study generalizations of the slope heuristics method : for robustness, multivariate case, etc...
- implement the new methods,
- evaluate their predictive performances through a large-scale simulation study for different models: linear regression, break-point detection, mixture models, graphical models, etc...

Keywords :

model selection, high-dimensional statistics, prediction, variable selection, penalization, slope heuristics.

Profile :

Final-year student in a master's program (or equivalent) in statistics, data sciences or in an engineering school with :

- strong knowledge in statistics, particularly in high-dimensional statistics,
- knowledge about sparse linear modeling and/or mixture models is an advantage but it is not a requirement,
- advanced **R** programming.

Practical details :

- Internship location : The internship will be co-supervised by Vincent Brault and Perrine Lacroix, and will take place at Laboratoire de Mathématiques Jean Leray (Nantes), with the possibility of visits or a short stay at Laboratoire Jean Kuntzmann (Grenoble).
- Internship period : 6-month, spring and summer 2026. The internship will start as soon as the master's program ends. Please note that the LMJL will be closed from July 24 to August 16, 2026; and that the student can take vacations in conformity with the legal rules.
- Internship remuneration according to the official civil service pay scale : approximately 650 euros net/month [3]. This internship is funded by the PULSAR Academy, the academy for young researchers in the Pays de La Loire region.
- To apply : Interested candidates can send a CV and a motivation [4] letter to Vincent Brault and Perrine Lacroix.

[3] Average depending on the number of days actually worked in the month

[4] Any application generated using a large language model may be rejected immediately before reaching the supervisors.

References

Akaike, H. (1973). Maximum likelihood identification of gaussian autoregressive moving average models. Biometrika, 60(2):255–265.

Baudry, J.-P., Maugis, C., and Michel, B. (2012). Slope heuristics: overview and implementation. Statistics and Computing, 22(2):455–470.

Birgé, L. and Massart, P. (2007). Minimal penalties for gaussian model selection. Probability theory and related fields, 138(1):33–73.

Brault, V., Baudry, J.-P., Maugis-Rabusseau, C., and Michel, B. (2012). Capushe: package de sélection de modèle. In 1ères Rencontres R.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society: series B (methodological), 39(1):1–22.

Droesbeke, J.-J., Fichet, B., and Tassi, P. (1989). Séries chronologiques: théorie et pratique des modèles ARIMA, volume 2. Economica.

Lacroix, P. (2022). Contributions to variable selection in high-dimension and its uses in biology. PhD thesis, Université Paris-Saclay.

Schwarz, G. (1978). Estimating the dimension of a model. The annals of statistics, pages 461–464.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology, 58(1):267–288.