# Sélection de modèles

Sylvain Arlot

Journées MAS 2010, Bordeaux

# Sélection de modèles

Session organisée par **Sylvain Arlot**

Le problème de la sélection de modèle consiste à choisir parmi une famille de modèles celui qui conduit à un estimateur de risque minimal, à l'aide des données uniquement. Le choix optimal étant appelé l'oracle, on cherche à prouver une *inégalité-oracle*, c'est-à-dire que le risque de l'estimateur sélectionné est inférieur à celui de l'oracle multiplié par une (petite) constante $C$ avec grande probabilité.

La *pénalisation* a souvent été utilisée pour cela avec succès, depuis les travaux de Mallows [4] et Akaike [1] notamment, puis ceux de Barron, Birgé et Massart [2, 3] dans un cadre non-asymptotique. Elle consiste à choisir le modèle qui minimise la somme du risque empirique sur le modèle et d'une pénalité mesurant la complexité du modèle. Le choix d'une pénalité adéquate est crucial pour optimiser le risque de l'estimateur final, à la fois du point de vue théorique et du point de vue pratique.

Cette session abordera, dans différents cadres statistiques (régression, estimation de densité, clustering, etc.), différentes questions liées au choix d'une pénalité optimale, sous un angle à la fois théorique et pratique :
- quelle est la *forme d'une pénalité adaptée* au problème de sélection de modèles posé ?
- comment *calibrer au mieux (à l'aide des données uniquement) les constantes* intervenant dans la pénalité (en particulier, via l'heuristique de pente proposée par Birgé et Massart [3]) ?
- comment *calculer de manière effective* la pénalité et le modèle sélectionné ?

*Références :*

[1] Akaike, Hirotugu. *Information theory and an extension of the maximum likelihood principle* (1973) In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, 267–281, Akadémiai Kiadó, Budapest.

[2] Barron, Andrew, Birgé, Lucien and Massart, Pascal. *Risk bounds for model selection via penalization* (1999) Probab. Theory Related Fields **113**, 301–413.

[3] Birgé, Lucien and Massart, Pascal. *Minimal penalties for Gaussian model selection* (2007) Probab. Theory Related Fields **138**, 33–73.

[4] Mallows, Colin L. *Some comments on $C_p$* (1973) Technometrics **15**, 661–675.

*Adresse de l'organisateur :*

Journées MAS 2010, Bordeaux

Sylvain Arlot
CNRS, Équipe-Projet Willow
Laboratoire d'Informatique de l'École Normale Supérieure (CNRS/ENS/INRIA, UMR 8548)
INRIA, 23 avenue d'Italie, CS 81321
75214 PARIS Cedex 13 - France
E-mail : sylvain.arlot@ens.fr
<http://www.di.ens.fr/~arlot/>

Session : Sélection de modèles

Journées MAS 2010, Bordeaux

Session : Sélection de modèles

## Estimation adaptative par sélection de partitions en rectangles dyadiques

par **Nathalie Akakpo** et Claire Lacour

Supposons que l'on veuille estimer une fonction $s$ définie sur le carré unité en se basant sur l'observation de $n$ variables indépendantes. Nous proposons une procédure basée sur une collection particulière de partitions du carré unité, composées chacune de rectangles dyadiques de dimensions éventuellement différentes. Sur chaque partition, nous définissons un estimateur polynomial par morceaux adéquat. Puis nous sélectionnons la meilleure partition à l'aide d'un critère de type moindres carrés pénalisé basé sur les données. Dans cet exposé, nous nous intéresserons à l'estimation de densité ou de densité conditionnelle. Nous proposons dans ce cadre une pénalité permettant d'obtenir une inégalité de type oracle. Notre estimateur possède alors également des propriétés d'adaptation au sens minimax, à constante près, sur des classes de fonctions de régularité éventuellement non homogène et non isotrope. De plus, il peut être implémenté avec une complexité algorithmique seulement linéaire en la taille de l'échantillon.

*Adresses :*

Nathalie Akakpo
Département de Mathématiques, Université Paris-sud ; Université Paris Descartes
Bât. 430
91405 Orsay France
E-mail : nathalie.akakpo@math.u-psud.fr
<http://www.math.u-psud.fr/~akakpo/>

Claire Lacour
Département de Mathématiques, Université Paris-Sud
Bât. 430
91405 Orsay France
E-mail : claire.lacour@math.u-psud.fr
<http://www.math.u-psud.fr/~lacour/>

Journées MAS 2010, Bordeaux

Session : Sélection de modèles

## Heuristique de pente en sélection de modèles pour des M-estimateurs à contraste régulier

par **Adrien Saumard**

Les procédures de sélection de modèles sont sensibles au choix des constantes dans les pénalités, choix qui se révèle souvent peu fondé en pratique, une sous-pénalisation pouvant dégrader considérablement la performance de l'algorithme associé. Birgé et Massart (2007) ont ainsi récemment introduit une méthode de calibration automatique des pénalités, appelée *heuristique de pente,* dont le but intrinsèque - contrairement à d'autres méthodes de calibration - est d'améliorer la performance en prédiction des algorithmes. Cette méthode se base en pratique sur un saut identifiable dans les dimensions des modèles sélectionnés, ce saut étant localisé autour d'un certain seuil de pénalisation appelé pénalité minimale. L'heuristique stipule alors que la pénalité optimale, qui sélectionne un estimateur dont le risque est équivalent à celui de l'oracle, vaut deux fois la pénalité minimale.

Le but de l'exposé est de valider cette heuristique et de montrer l'optimalité non-asymptotique de l'estimateur sélectionné dans un cadre générique nouveau que nous définirons et que nous appellerons "$M$-estimation à contraste régulier". Dans ce cadre, nous retrouverons et généraliserons certains résultats de Arlot et Massart (2009), et Lerasle (2009). Nous validerons aussi pour la première fois l'heuristique de pente pour un risque non quadratique, dans le cas de l'estimation de la densité par maximum de vraisemblance.

*Adresse :*
Adrien Saumard
Université Rennes 1, IRMAR
UFR Mathématiques, Campus de Beaulieu
35042 Rennes France
E-mail : adrien.saumard@univ-rennes1.fr

Journées MAS 2010, Bordeaux

Session : Sélection de modèles

## Pratique de l'heuristique de pente et le package CAPUSHE

par **Jean-Patrick Baudry**, Cathy Maugis et Bertrand Michel

La mise en œuvre des méthodes "data-driven" de calibration de critères pénalisés, issues de l'heuristique de pente de Birgé et Massart (2007), implique des difficultés pratiques. Nous discutons et comparons les deux approches disponibles : le "saut de dimension" et l'"estimation directe de la pente". Nous présentons une solution pour la mise en œuvre de cette dernière approche, qui repose sur une étude de la stabilité du modèle sélectionné. Les solutions proposées sont implémentées dans le package CAPUSHE qui permet une application simple et conviviale de ces méthodes.

*Adresses :*

Jean-Patrick BAUDRY
Université Paris-sud ; INRIA, Projet SELECT ; MAP5, Université Paris Descartes
45 rue des Saint Pères
75270 Paris Cedex 06, France
E-mail : Jean-Patrick.Baudry@math.u-psud.fr
<http://www.math-info.univ-paris5.fr/~baudryjp/>

Cathy MAUGIS
Institut de Mathématiques de Toulouse
135, avenue de Rangueil
31077 Toulouse Cedex 4, France
E-mail : cathy.maugis@insa-toulouse.fr
<http://www.math.univ-toulouse.fr/~maugis/>

Bertrand MICHEL
LSTA, Université Pierre et Marie Curie
175 rue du Chevaleret
75013 Paris, France
E-mail : bertrand.michel@upmc.fr
<http://www.lsta.upmc.fr/michelb.html>

Journées MAS 2010, Bordeaux

Session : Sélection de modèles

## Clustering et sélection de variables sur des données génétiques

par **Dominique Bontemps** et Wilson Toussile

Nous nous intéressons au problème d'estimer les variables pertinentes et le nombre de composantes d'une loi de mélange pour des données génotypiques multilocus. Un critère du maximum de vraisemblance pénalisé est proposé, et une inégalité oracle non-asymptotique est obtenue. En outre, sous des conditions faibles portant sur la distribution qui a généré les observations, le modèle sélectionné est asymptotiquement consistent. D'un point de vue pratique, la pénalité est définie à une constante multiplicative près, et celle-ci est calibrée par l'heuristique de pente. Sur des données simulées la procédure de sélection fait mieux que des critères classiques tels que BIC et AIC. Le nouveau critère apporte une réponse à la question : "Quel critère choisir en fonction de la taille de l'échantillon ?".

*Adresses :*

Dominique Bontemps
Département de Mathématiques
Univ. Paris-Sud 11
Bât. 430
91405 Orsay France
E-mail : dominique.bontemps@math.u-psud.fr
<http://www.math.u-psud.fr/~bontemps/>

Wilson Toussile
Département de Mathématiques
Univ. Paris-Sud 11
Bât. 430
91405 Orsay France
E-mail : Wilson.Toussile@math.u-psud.fr
<http://www.math.u-psud.fr/~toussile/>