Lois à priori parcimonieuses et estimation en grande dimension

Pierre Alquier

Journées MAS 2010, Bordeaux

Lois à priori parcimonieuses et estimation en grande dimension

Session organisée par **Pierre Alquier**

Dans le but d'obtenir des résultats théoriques (inégalités oracle) et de bonnes performances pratiques dans le contexte de l'estimation en grande dimension (et de la sélection de modèle), il est nécessaire de controler la complexité des estimateurs proposés. Dans les travaux PAC-Bayésiens (Catoni [1], Dalalyan et Tsybakov [2], ...) ainsi que Bayésiens (par exemple Ghosal, Lember et van der Vaart [3]), une loi π a priori sur le paramètre permet ce contrôle.

Considérons par exemple la régression linéaire en grande dimension : $y \sim \mathcal{N}(X\beta^*, \sigma^2 I_n)$ pour $\beta \in \mathbf{R}^p$ avec $p > n$. Les méthodes de moindres carrés penalisés, pour $\lambda \geq 0$, $\gamma \geq 0$,

$$\min_{\beta} \left\{ \|y - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} |\beta|^\gamma \right\},$$

qui incluent les pénalisations de type AIC ou BIC (pour $\gamma = 0$), le LASSO (Tibshirani [4], pour $\gamma = 1$) et la *Ridge Regression* (pour $\gamma = 2$) peuvent etre vues comme des maximum a posteriori d'estimateurs bayésiens avec comme loi a priori $\pi(d\beta) \propto \exp(-\lambda \sum_{j=1}^{p} |\beta|^\gamma) d\beta$. Cet exemple illustre l'importance du choix de π sur les propriétés de l'estimateur obtenu : implémentable, ou non, pour de grandes valeurs de p ; possibilité d'estimer correctement le support du paramètre ou non, etc...

L'objectif de cette session est d'illustrer l'importance de π dans les propriétés théoriques de l'estimateur obtenu, ainsi que dans ses performances pratiques dans des applications.

Références :

[1] Catoni, O., *PAC-Bayesian Supervised Classification : The Thermodynamics of Statistical Learning*, IMS Lecture Notes, vol. 56, 2008.

[2] Dalalyan, A. & Tsybakov, A., Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity, *Machine Learning*, 72, pp 39-61, 2008.

[3] Ghosal, S., Lember, J. & van der Vaart, A. W., Nonparametric Bayesian Model Selection and Averaging, *Electronic Journal of Statistics*, 2, pp 63-89, 2008.

[4] Tibshirani, R., Regression Shrinkage and Selection via the LASSO. JRSS B., 58, pp 267-288, 1996.

Adresse de l'organisateur :

Journées MAS 2010, Bordeaux

Pierre Alquier
Laboratoire de Probabilités et Modèles Aléatoires, Univ. Paris 7
175, rue du Chevaleret
75013 Paris
E-mail : alquier@math.jussieu.fr
<http://alquier.ensae.net/>

Session : Lois à priori parcimonieuses et estimation en grande dimension

Journées MAS 2010, Bordeaux

Session : Lois à priori parcimonieuses et estimation en grande dimension

PAC-Bayesian Inequalities and Robust Estimation

par **Olivier Catoni** et Jean-Yves Audibert

In this talk, we will outline the main ideas underlying PAC-Bayesian bounds. We will show connections with empirical process theory and relate the bounds to information theoretic measures of the complexity of models and estimators. We will also explain how to combine the PAC-Bayesian approach with some robust exponential inequalities derived from the introduction of some special family of influence functions. We will illustrate this presentation with the analysis of least square regression with empirical design. We will present an improved analysis of the ordinary least square estimator and introduce new thresholded estimators with finite sample exponentially consistent bounds under weak polynomial moment assumptions, allowing the Gram matrix to be ill-conditioned.

Adresses :

Olivier Catoni
CNRS
Département de Mathématiques et Applications, Ecole Normale Supérieure
45 rue d'Ulm
75230 Paris CEDEX 05
E-mail : olivier.catoni@ens.fr
<http://www.dma.ens.fr/~catoni/homepage/newpage.html>

Jean-Yves Audibert
CERTIS
Ecole Nationale des Ponts et Chaussées
6, avenue Blaise Pascal - Cité Descartes
Champs-sur-Marne
77455 Marne-la-Vallée CEDEX 2
E-mail : audibert@imagine.enpc.fr
<http://certis.enpc.fr/~audibert/>

Journées MAS 2010, Bordeaux

Session : Lois à priori parcimonieuses et estimation en grande dimension

Sparsity oracle inequalities for mirror avaraging aggregate

par **Arnak Dalalyan** et Alexandre Tsybakov

We consider the problem of aggregating the elements of a (possibly infinite) dictionary for building a decision procedure, that aims at minimizing a given criterion. Along with the dictionary, an independent identically distributed training sample is assumed available on which the performance of a given procedure can be tested. In a fairly general set-up, we establish an oracle inequality for the Mirror Averaging aggregate based on any prior distribution. This oracle inequality is applied in the context of sparse coding for different tasks of statistics and machine learning such as regression, density estimation and binary classification.

Adresses :

Arnak Dalalyan
IMAGINE / CERTIS, Ecole des Ponts - ParisTech
6, Av Blaise Pascal - CitÉdescartes
Champs-sur-Marne
77455 Marne-la-Vallé CEDEX 2
E-mail : dalalyan@certis.enpc.fr
<http://certis.enpc.fr/~dalalyan/>

Alexandre Tsybakov
CREST, Laboratoire de Statistique, et Université Paris 6, LPMA
CREST-LS, Timbre J340
3, avenue Pierre Larousse
92240 Malakoff CEDEX email
<http://www.proba.jussieu.fr/~tsybakov/tsybakov.html>

Journées MAS 2010, Bordeaux

Session : Lois à priori parcimonieuses et estimation en grande dimension

Modélisation de pannes sur un réseau électrique souterrain

par **Sophie Donnet** et Judith Rousseau

Nous nous intéressons à la modélisation des pannes sur un réseau électrique souterrain, ce réseau étant composé de deux types de matériels : câbles et accessoires. Après une panne sur l'un ou l'autre des matériels, la partie endommagée est retirée et remplacée par un ou deux accessoires (selon que la panne a lieu sur un accessoire ou sur le câble lui-même). Ainsi la structure du réseau est modifiée au cours du temps. La modélisation proposée vise à prendre en compte l'évolution temporelle du réseau, et en particulier l'évolution du nombre d'accessoires dans le réseau dans le but d'estimer les taux de panne des différentes composantes du réseau. Afin de ne pas s'appuyer sur les relevés des types de pannes (câble ou accessoire), nous supposons les causes des incidents inconnues. Pour ce faire, nous proposons un mod'ele basé sur un processus de Poisson. Pour estimer les param'etres impliqués dans la modélisation des pannes, nous considérons une approche bayésienne. La loi a posteriori est obtenue par un algorithme de Gibbs. Cependant une première étude sur données simulées a montré l'inluence cruciale du nombre d'accessoires présents sur le réseau au début de l'étude. Ce nombre initial est inconnu dans la pratique et doit être estimé. Dans ce travail, nous proposons de construire une loi a priori sur ce nombre initial reposant sur le comportement asymptotique du processus.

Adresses :
Sophie Donnet
CEREMADE, Université Paris Dauphine
Place du Maréchal De Lattre De Tassigny
75775 Paris CEDEX 16
E-mail : donnet@ceremade.dauphine.fr
<http://www.ceremade.dauphine.fr/~donnet/>

Journées MAS 2010, Bordeaux

Judith Rousseau
CREST et CEREMADE (Dauphine)
Université Paris Dauphine
Place du Maréchal De Lattre De Tassigny
75775 Paris CEDEX 16
E-mail : rousseau@ceremade.dauphine.fr
<http://www.ceremade.dauphine.fr/~rousseau/>

Session : Lois à priori parcimonieuses et estimation en grande dimension

Journées MAS 2010, Bordeaux

Session : Lois à priori parcimonieuses et estimation en grande dimension

PAC-Bayesian approach for kernel methods

par **Joseph Salmon** et Erwan Le Pennec

In this work on regression with Gaussian error, we study an agregation procedure relying on the exponential weigthing scheme described in Dalalyan and Tsybakov [1]. We obtain PAC-Bayes oracle inequalities in this context valid in both the fixed design case and the random design case. These inequalities are obtained by techniques derived from those described in Catoni [2] and Audibert [3]. We apply those results to the selection of an "optimal" window for Nadaraya-Watson type estimators and obtain a provably efficient estimator implemented with a MCMC-type algorithm similar to the one proposed by Dalalyan and Tsybakov [3].

Références :

[1] A. Dalalyan and A. Tsybakov, Sparse regression learning by aggregation and Langevin Monte-Carlo, in 22th Annual Conference on Learning Theory, COLT, 2009.

[2] O. Catoni, Statistical learning theory and stochastic optimization, ser. Lecture Notes in Mathematics. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, 2001.

[3] J.-Y. Audibert, Aggregated estimators and empirical complexity for least square regression, Ann. Inst. H. Poincaré Probab. Statist., vol. 40, no. 6, pp. 685-736, 2004.

Adresses :

Joseph Salmon
Laboratoire de Probabilités et Modèles Aléatoires, Univ. Paris 7
175, rue du Chevaleret
75013 Paris
E-mail : salmon@math.jussieu.fr
<http://people.math.jussieu.fr/~salmon/>

Erwan Le Pennec
Laboratoire de Probabilités et Modèles Aléatoires, Univ. Paris 7
Projet SELECT / INRIA Saclay / Université Paris Sud
LPMA
175, rue du Chevaleret
75013 Paris
E-mail : salmon@math.jussieu.fr
<http://people.math.jussieu.fr/~salmon/>