



INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team select

Model selection in statistical learning

Saclay - Île-de-France

Theme : Optimization, Learning and Statistical Methods

Activity
R *report*

2010

Table of contents

1. Team	1
2. Overall Objectives	1
2.1. Model selection in Statistics	1
2.2. Highlights	2
3. Scientific Foundations	2
3.1. General presentation	2
3.2. A non asymptotic view for model selection	2
3.3. Taking into account the modelling purpose in model selection	2
3.4. Bayesian model selection	2
3.5. Nonlinear mixed effect models	3
4. Application Domains	3
4.1. Introduction	3
4.2. Curves classification	3
4.3. Computer Experiments and Reliability	3
4.4. Neuroimaging	3
4.5. Population pharmacology	3
5. Software	4
5.1. MIXMOD software	4
5.2. MONOLIX software	4
6. New Results	5
6.1. Model selection in Regression and Classification	5
6.2. Statistical learning methodology and theory	7
6.3. Reliability and Computer Experiments	7
6.4. Curves classification, denoising and forecasting	8
6.5. Neuroimaging, Statistical analysis of fMRI data	8
6.6. Nonlinear mixed effects model	9
7. Contracts and Grants with Industry	9
7.1. Contracts with EDF	9
7.2. The Monolix software project	9
7.3. Other contracts	10
7.4. Project GAS	10
7.5. Project EHPOC	10
8. Other Grants and Activities	10
8.1. National Actions	10
8.2. European actions	11
8.3. International Initiatives	11
9. Dissemination	11
9.1. Scientific Community animation	11
9.1.1. Editorial responsibilities	11
9.1.2. Invited conferences	11
9.1.3. Scientific animation	11
9.2. Teaching	12
10. Bibliography	12

1. Team

Research Scientists

Gilles Celeux [Team Vice-Leader, Senior Researcher INRIA]
Marc Lavielle [Senior Researcher INRIA detached from Université Paris 5]
Erwan Le Pennec [Junior Researcher INRIA]
Kevin Bleakey [Junior Researcher INRIA since October 2010]

Faculty Members

Pascal Massart [Team Leader, Professor Université Paris-Sud]
Christine Keribin [Associate Professor]
Marie-Anne Poursat [Associate Professor]
Jean-Michel Poggi [Professor Université Paris 5]

External Collaborator

Yves Auffray [Dassault]

Technical Staff

Benoît Charles
Kaelig Chatel
Rémi Fouchereau [until August 31, 2010]
Morgan Guery
Hector Mesa
ERic Blaudez
Jean-François Si Abdallah

PhD Students

Pierre Barbillon [MESR grant]
Jean-Patrick Baudry [ENS grant]
Pierre Connault [CIFRE grant]
Caroline Meynet [MESR grant]
Jairo Cugliari-Duhalde [CIFRE grant]
Maud Delattre [Région IDF grant]
Mohammed El Anbari [France-Marocco grant]
Shuai Fu [EDF-INRIA Cifre grant]
Robin Genuer [MESR grant]
Merlin Keller [CEA-INRIA grant]
Cyprien Mbogning [INRIA grant]
Vincent Michel [INRIA grant]
Wilson Toussile [France-Cameroun grant]

Post-Doctoral Fellow

Andrea Rau [since October 1, 2010]

Administrative Assistant

Katia Evrat [TR partially]

2. Overall Objectives

2.1. Model selection in Statistics

The research domain for the SELECT project is statistics. Statistical methodology has made great progress over the past few decades, with a variety of statistical learning software packages that support many different methods and algorithms. Users now face the problem of choosing among them, to select the most appropriate method for their data sets and objectives. The problem of model selection is an important but difficult problem both theoretically and practically. Classical model selection criteria, which use penalized minimum-contrast criteria with fixed penalties, are often based on unrealistic assumptions.

SELECT aims to provide efficient model selection criteria with data-driven penalty terms. In this context, SELECT expects to improve the toolkit of statistical model selection criteria from both theoretical and practical perspectives. Currently, SELECT is focusing its effort on variable selection in statistical learning, non-linear regression models with random effects, hidden-structure models and supervised classification. Its domains of application concern reliability, curves classification, phylogeny analysis and classification in genetics. New developments of SELECT activities are concerned with applications in biostatistics (statistical analysis of fMRI data, population pharmacology) and population genetics.

2.2. Highlights

Pascal Massart has been nominated senior member of the Institut Universitaire de France.

3. Scientific Foundations

3.1. General presentation

We learned from the applications we treated that some assumptions which are currently used in asymptotic theory for model selection are often irrelevant in practice. For instance, it is not realistic to assume that the target belongs to the family of models in competition. Moreover, in many situations, it is useful to make the size of the model depend on the sample size which make the asymptotic analysis breakdown. An important aim of SELECT is to propose model selection criteria which take these practical constraints into account.

3.2. A non asymptotic view for model selection

An important purpose of SELECT is to build and analyze penalized log-likelihood model selection criteria that are efficient when the number of models in competition grows to infinity with the number of observations. Concentration inequalities are a key tool for that purpose and lead to data-driven penalty choice strategies. A major issue of SELECT consists of deepening the analysis of data-driven penalties both from the theoretical and the practical side. There is no universal way of calibrating penalties but there are several different general ideas that we want to develop, including heuristics derived from the Gaussian theory, special strategies for variable selection and using resampling methods.

3.3. Taking into account the modelling purpose in model selection

Choosing a model is not only difficult theoretically. From a practical point of view, it is important to design model selection criteria that accommodate situations in which the data probability distribution P is unknown and which take the model user's purpose into account. Most standard model selection criteria assume that P belongs to one of a set of models, without considering the purpose of the model. By also considering the model user's purpose, we avoid or overcome certain theoretical difficulties and can produce flexible model selection criteria with data-driven penalties. The latter is useful in supervised Classification and hidden-structure models.

3.4. Bayesian model selection

The Bayesian approach to statistical problems is fundamentally probabilistic. A joint probability distribution is used to describe the relationships among all the unknowns and the data. Inference is then based on the posterior distribution i.e. the conditional probability distribution of the parameters given the observed data. Exploiting the internal consistency of the probability framework, the posterior distribution extracts the relevant information in the data and provides a complete and coherent summary of post-data uncertainty. Using the posterior to solve specific inference and decision problems is then straightforward, at least in principle.

3.5. Nonlinear mixed effect models

Mathematical modelling of the dynamic processes involved in biological processes constitutes an important application in biostatistics. Mixed effect models are very useful for modelling the variability within a population of these dynamic processes. Several statistical issues can be studied related to these models, such as parameter estimation, model selection (covariate model through the specification of fixed effect structure, covariance model for random effects), models defined by Ordinary or Stochastic Differential Equations, left censored models, hidden Markov models as well as design optimization for the trial itself.

4. Application Domains

4.1. Introduction

A key goal of SELECT is to produce methodological contributions in statistics. For this reason, the SELECT team works with applications that serve as an important source of interesting practical problems and require innovative methodologies to address them. Most of our applications involve contracts with industrial partners, e.g. in reliability and pharmacology, although we also have several more academic collaborations, e.g. genomics, genetics and neuroimaging.

4.2. Curves classification

The field of classification for complex data as curves, functions, spectra and time series is important. Standard data analysis questions are being revisited to define new strategies that take the functional nature of the data into account. Functional data analysis addresses a variety of applied problems, including longitudinal studies, analysis of fMRI data and spectral calibration.

We are focusing on unsupervised classification. In addition to standard questions as the choice of the number of clusters, the norm for measuring the distance between two observations, and the vectors for representing clusters, we must also address a major computational problem. The functional nature of the data needs to be design efficient anytime algorithms.

4.3. Computer Experiments and Reliability

Since several years, SELECT has collaborations with EDF-DER *Maintenance des Risques Industriels* group. An important theme concerns the resolution of inverse problems using simulation tools to analyse uncertainty in highly complex physical systems. A collaboration on an analogous topic is developed with Dassault Aviation.

The other major theme concerns probabilistic modelling in fatigue analysis in the context of a research collaboration with SAFRAN an high-technology group (Aerospace propulsion, Aircraft equipment, Defense Security, Communications).

4.4. Neuroimaging

Since 2007 SELECT participates to a working group with team Neurospin (CEA-INSERM-INRIA) on Classification, Statistics and fMRI (functional Magnetic Resonance Imaging) analysis. In this framework two theses are co-supervised by SELECT and Neurospin researchers (Merlin Keller since October 2006 and Vincent Michel since October 2007). The aim of this research is to determine which parts of the brain are activated by different types of stimuli. A model selection approach is useful to avoid "false-positive" detections.

4.5. Population pharmacology

Pharmacokinetic (PK) and pharmacodynamic (PD) studies (studies investigating the dose-concentration and concentration-effect relationships of drugs) show for many drugs a large variability of pharmacokinetic and pharmacodynamic parameters between individuals. Pharmacokinetic parameters describe processes such as absorption, diffusion and metabolism of drugs. The so-called "population PK/PD approach" has been developed to characterize and quantify this variability. We have developed a complete methodology for the analysis of PK/PD data using a maximum likelihood approach.

An important application is the study of anti-HIV treatment. The efficiency of antiretroviral treatments, whether in HIV or hepatitis B or C pathologies, is quantified by the decrease in viral loads. Models have been developed to describe the time-course of this decrease through a system of ODE, taking into account the physiology of viral replication and the action mechanisms of the different therapeutic options. There is a large inter-patient variability in these pathologies, and the joint study of viral load decrease through mixed effect models in a set of patients provides a better understanding of differences in the response to treatment.

5. Software

5.1. MIXMOD software

Participants: Gilles Celeux [Correspondant], Jean-François Si Abdallah.

MIXMOD is being developed in collaboration with Christophe Biernacki, Florent Langrognet (Université de Franche-Comté) and Gérard Govaert (Université de Technologie de Compiègne). MIXMOD (MIXture MODelling) software fits mixture models to a given data set with either a clustering or a discriminant analysis purpose. MIXMOD uses a large variety of algorithms to estimate mixture parameters, e.g., EM, Classification EM, and Stochastic EM. They can be combined to create different strategies that lead to a sensible maximum of the likelihood (or completed likelihood) function. Moreover, different information criteria for choosing a parsimonious model, e.g. the number of mixture component, some of them favoring either a cluster analysis or a discriminant analysis view point, are included. Many Gaussian models for continuous variables and multinomial models for discrete variable are available. Written in C++, MIXMOD is interfaced with SCILAB and MATLAB. The software, the statistical documentation and also the user guide are available on the Internet at the following address: <http://www-math.univ-fcomte.fr/mixmod/index.php>.

Since November 2008, a new expert engineer Jean-François Si Abdallah has been hired for two years to continue to enrich the software, improve the performances and code a proper graphical user interface in MIXMOD which Version 1 will be available at the end of the year 2010.

5.2. MONOLIX software

Participants: Marc Lavielle [Correspondant], Benoît Charles, Kaelig Chatel, Eric Blaudez, Hector Mesa.

MONOLIX (<http://software.monolix.org>) is free software dedicated to the analysis of non linear mixed effects models. The objective of the MONOLIX software is to perform:

- Parameter estimation (computing the maximum likelihood estimator of the parameters, without any approximation of the model, computing standard errors for the maximum likelihood estimator),
- Model selection (comparing several models using some information criteria (AIC, BIC), testing hypotheses using the Likelihood Ratio Test, testing parameters using the Wald Test),
- Goodness of fit plots,
- Data simulation.

Several stochastic algorithms are used in MONOLIX: Stochastic approximation of EM (SAEM), Importance Sampling, MCMC, and Simulated Annealing... Theoretical properties of the proposed algorithms and practical applications were published in several papers.

Marc Lavielle has presented the software in several occasions:

- PAGE meeting, Berlin, June 2010,
- Sheffield University (UK), January 2010,
- Sanofi-Aventis (France), October 2010,

Version 3.2 of MONOLIX is available since November 2010. This version of the software was developed thanks to the financial support of Novartis, Roche, J&J, Sanofi-Aventis, Astrazeneca, Exprimio. This version was presented during the MONOLIX Day on November 26th at the Hotel Holliday Inn - Canal de la Villette, Paris.

The MONOLIX Project consists primarily of developing the next versions of the MONOLIX software with a view to raising its level of functionalities and responding to major requirements of the bio-pharmaceutical industry.

The MONOLIX Project is carried out by INRIA, and sponsored by the Industry.

The MONOLIX Scientific Guidance Committee involves representatives of the sponsors.

We have obtained from INRIA Saclay-Île-de-France an ADT (Action Développement Logiciel) to hire two engineers (Benoît Charles until June 2011, Hector Mesa until December 2010).

We have obtained from DIGITEO an OMTE (Opération de Maturation Technico-Économique) to hire one engineer (Kaelig Chatel from September 2009 to March 2011), market assessment and intellectual property coaching.

6. New Results

6.1. Model selection in Regression and Classification

Participants: Jean-Patrick Baudry, Gilles Celeux, Mohammed El Anbari, Robin Genuer, Erwan Le Pennec, Pascal Massart, Caroline Meynet, Jean-Michel Poggi.

In collaboration with Serge Cohen (IPANEMA Soleil), Erwan Le Pennec has proposed a novel way to use spatial information for hyperspectral image segmentation. They rely on a Gaussian mixture model in which the proportions are piece-wise constant in the spatial domain. They proposed a penalization scheme to select simultaneously the number of clusters, the segmentation as well as the type of covariance structure. This scheme have been implemented within MIXMOD and used with success with conservation sciences data. Their algorithm is supported by a theoretical work on conditional density estimation. They have derive oracle inequalities relating a tensorized Kullback-like divergence of the estimation error with a deterministic quantity measuring a penalized tensorial Kullback divergence approximation property of the model collection. They obtain those results for their spatial Gaussian mixture model as well as for more classical histogram type conditional density models. The theoretical penalty is known up to a factor that has to be determined numerically. Serge Cohen and Erwan Le Pennec are investigating a slope heuristic approach to handle this issue. They want next to tackle the issue of dimension reduction in an unsupervised segmentation context [64].

In collaboration with Marie-Laure Martin-Magniette (URGV et UMR AgroParisTech/INRA MIA 518) and Cathy Maugis (INSA Toulouse) has extended their variable selection procedure for model-based clustering to supervised classification. Identifiability and consistence of their model has been proved. Their forward selection procedure lead to dramatically improved error rates for quadratic discriminant analysis [71]. These variable selection procedures are in particular used for genomics applications which is the result of a collaboration with researchers of of URGV (Evry Genopole).

Jean-Patrick Baudry in collaboration with Cathy Maugis (INSA Toulouse) and Bertrand Michel (Université Paris 6) [61] have developed a Matlab package for the use of the slope heuristics of Birgé and Massart (dimension jump, slope estimation, ...). The aim is twofold: first to propose solutions to overcome the practical difficulties involved by its practical application and second to provide a user-friendly solution to apply the slope heuristics. They have submitted an overview about the slope heuristics to introduce this package [60].

Jean-Patrick Baudry and Gilles Celeux, in collaboration with Ana Maria Ferreira and Margarida Cardoso (Lisbon University), proposed a model selection criterion which can be helpful when it can be interesting to find a solution well-related to an external classification available *a priori*. The development of the methodology and applications to real data in various fields, and notably in the professional development field, are in progress.

Pascal Massart and Caroline Meynet have analyzed the performance of the Lasso as regards its ℓ_1 -regularization properties in a general Gaussian framework which includes the fixed design regression or the white noise frameworks. They have provided an ℓ_1 -oracle inequality satisfied by the Lasso in the case of a finite dictionary. This result does not require any assumption and proves that the Lasso performs almost as well as the deterministic Lasso, provided that the regularization parameter is properly chosen. They proposed a new estimator to deal with infinite countable dictionaries. It is an ℓ_0 -penalized estimator among a sequence of Lasso estimators associated to a dyadic sequence of growing truncated dictionaries. They provided an oracle inequality satisfied by this estimator and proved that it performs as well as greedy algorithms with the advantage of being adaptive [23].

In collaboration with Professor Abdallah Mkhadri (University of Marrakesh, Morocco), Gilles Celeux supervised the thesis of Mohammed El Anbari which concern regularisation methods in linear regression. This year, in collaboration with Professor Abdallah Mkhadri (University of Marrakesh, Morocco), Mohammed El Anbari proposed a method to simultaneously select variables and favor a grouping effect where strongly correlated predictors tend to be in or out of the model together. Numerical experiments showed that their method can be preferred to Elastic-Net when the number of variables is less or equal to the sample size and remain competitive otherwise. Moreover, they have proposed AdaGril an extension of the adaptive Elastic Net which incorporates information redundancy among correlated variables for model selection and estimation. Under weak conditions, They have established an oracle property of AdaGril. Numerical experiments show in some cases of AdaGril outperforms its competitors.

In collaboration with Jean-Michel Marin (Université de Montpellier) and Christian P. Robert (CEREMADE, Université Paris Dauphine) Gilles Celeux and Mohammed El Anbari highlight the interest of Bayesian regularization methods, using hierarchical non informative priors, compared with standard regularisation methods in a poorly informative context through numerical experiments [55].

Jean-Michel Poggi is the supervisor of the PhD Thesis of Robin Genuer since September 2007 dedicated to Random Forests and related algorithms for variable selection in regression or classification. Random Forest, due to Leo Breiman in 2001, proceeds by aggregation decision trees according to two random perturbations. The first one perturbs the learning sample according to the bootstrap principle and the second one acts on the covariate space by choosing randomly a small number of explanatory variables to split a tree node. Surprisingly, this algorithm is extremely powerful for regression and classification problems, not only for prediction but also for variable selection purposes. The PhD thesis, obtained in november 2010, is articulated following three directions:

- The theoretical direction concerns mathematical understanding of the reasons of this amazing behaviour. A simplified version of random forests, called purely random forests, which can be theoretically handled more easily, has been considered. In the regression context with a one-dimensional predictor space, both random trees and random forests has been proved to reach minimax rate of convergence. In addition, it is proven that compared to random trees, random forests improve accuracy by reducing the estimator variance by a factor of three fourths [57].
- The second methodological direction aims at improving the knowledge about how to tune the parameters and to propose, with Jean-Michel Poggi and Christine Tuleau-Malot, a new variable selection strategy [17]. It includes computer intensive simulations and comparisons based on well-known real data sets. Two classical issues of variable selection are considered. The first one is to find important variables for interpretation and the second one aims to design a good parsimonious prediction model. The main contribution is twofold: to provide some experimental insights about the behavior of the variable importance index based on random forests and to propose a strategy involving a ranking of explanatory variables using the random forests score of importance and a stepwise ascending variable introduction strategy.
- The last one is of applied nature and takes place on the joint working group between SELECT and Neurospin (INRIA, CEA) dedicated to statistical methods for fMRI new data in order to improve knowledge about brain activities. It aims to develop ad-hoc variable selection strategies (see [41]).

6.2. Statistical learning methodology and theory

Participants: Jean-Patrick Baudry, Gilles Celeux, Christine Keribin, Erwan Le Pennec, Pascal Massart, Jean-Michel Poggi.

Unsupervised segmentation is an issue similar to unsupervised classification with an added spatial aspect. Functional data is acquired on points in a spatial domain and the goal is to segment the domain in homogeneous domain. The range of applications includes hyperspectral images in conservation sciences, fMRI data and all spatialized functional data. Erwan Le Pennec is focusing on the questions of the way to handle the spatial component from both the theoretical and the practical point of views as well as the choice of the number of clusters. Furthermore, as functional data require heavy computation, he is required to propose numerically efficient algorithms.

After studying the state of the art on variational methods [32], Christine Keribin in collaboration with Gilles Celeux and Gérard Govaert (UTC Compiègne) has compared the behavior of the variational method with the use of the Stochastic EM algorithm, for the latent block model (case of biclustering, simultaneous classification of rows and columns of a matrix). This has been achieved on benchmark data and is leading to the use of the SEM algorithm to find an appropriate initialization for the VB algorithm [43]). This study will be continued on real data and extend to the construction of model selection methods to choose a relevant latent block model.

6.3. Reliability and Computer Experiments

Participants: Yves Auffray, Pierre Barbillon, Gilles Celeux, Pierre Connault, Rémi Fouchereau, Shuai Fu, Pascal Massart.

In the computer experiments field, the goal is to approximate an expensive black box function from a limited number of evaluations. The choice of these evaluations i.e. the choice of a design of (computer) experiments is a major issue.

have justified to take a design satisfying to the MAXIMIN criterion by using results from the approximation theory. In the case where the black box function is to be approximated on a hypercubic domain, the standard strategy consists of taking a MAXIMIN design within a class of Latin hypercube Designs. However, the Latin hypercube sampling is pointless in a non hypercubic domain. Yves Auffray, Pierre Barbillon and Jean-Michel Marin ([53]) proposed a simulated annealing algorithm, implemented in C, which aims at obtaining a MAXIMIN design in any bounded connected domain. They proved the convergence of their algorithm. This year Yves Auffray and Pierre Barbillon, in collaboration with Jean-Michel Marin (Université de Montpellier) have considered estimating the probability of rare events in the context of computer experiments. These rare events depends on the output of a physical model with random input variables. Since the model is only known through an expensive black box function, a crude Monte Carlo estimator does not perform well. Two strategies have been developed to cope with this difficulty: a Bayesian estimate and an importance sampling method. Both methods relies on Kriging metamodeling. They are able to achieve sharp upper confidence bounds on the rare event probability. These methods have been applied to a toy example and a real case study which consists of finding an upper bound of the probability that the trajectory of an airborne load collides the aircraft that has released it.

In the framework of a CIFRE convention with EDF (MRI departement), Shuai Fu has started a thesis supervised by Gilles Celeux. It concern the resolution of non linear inverse problems for the quantification of uncertainties in a physical model. Noisy observed data (Y) are dependent through a known but complex function H from non observed data X . The aim is to estimate parameters of the probability distribution of (X) and the variance of the noise. This problem has a missing data structure and has been solved with a method coupling the use of the Stochastic EM algorithm with a MCMC method and a Kriging approximation of the H function [10] under identifiability constraints. In order to overcome those limitations, Shuai Fu embedded the inverse problem in a Bayesian framework through an hybrid MCMC algorithm. Preliminary numerical experiments show the importance of choosing a smart design. Assessing the quality of the approximation and defining efficient adaptive designs are the next steps to be considered.

Moreover a complete numerical study has been achieved in the context of the project EHPOC by Rémi Fouchereau with the help of Pierre Barbillon, Gilles Celeux and Shuai Fu on Bayesian inference for inverse problems in industrial uncertainty analysis [54].

6.4. Curves classification, denoising and forecasting

Participant: Jean-Michel Poggi.

In collaboration with Farouk Mhamdi and Meriem Jaidane (ENIT, Tunis, Tunisia), Jean-Michel Poggi proposed, in [37], a method for trend extraction from seasonal time series through the Empirical Mode Decomposition (EMD). Experimental comparison of trend extraction based on EMD, X11, X12 and Hodrick Prescott filter are conducted. First results show the eligibility of the blind EMD trend extraction method. Tunisian real peak load is also used to illustrate the extraction of the intrinsic trend.

In collaboration with Mina Aminghafari (Amirkabir University, Teheran), Jean-Michel Poggi made uses of wavelets in a statistical forecasting purpose for time series. Recent approaches involve wavelet decompositions in order to handle non stationary time series. They study and extended an approach proposed by Renaud et al., to estimate the prediction equation by direct regression of the process on the Haar non-decimated wavelet coefficients depending on its past values. The new variants are used first for stationary data and after for stationary data contaminated by a deterministic trend [5].

Jean-Michel Poggi is the supervisor (with A. Antoniadis) of the PhD Thesis of Jairo Cugliari-Duhalde which takes place in a CIFRE convention with EDF. It is strongly related to the use of wavelets together with curves clustering in order to perform accurate load consumption forecasting. The thesis develops methodological and applied aspects linked to the electrical context as well as theoretical ones by introducing exogeneous variables in the context of nonparametric forecasting time series (see [33] and [36]).

6.5. Neuroimaging, Statistical analysis of fMRI data

Participants: Gilles Celeux, Robin Genuer, Merlin Keller, Christine Keribin, Vincent Michel, Jean-Michel Poggi.

This research takes place as part of a collaboration with Neurospin (<http://www.math.u-psud.fr/select/reunions/neurospin/Welcome.html>).

Vincent Michel began his PhD in October 2007 under the supervision of Gilles Celeux, Christine Keribin (Paris XI/ Select) and Bertrand Thirion (Parietal). His thesis [3] focused on the different approaches of statistical learning for studying the neural code of cognitive functions, based on brain functional Magnetic Resonance Imaging (fMRI) data. In particular, he has studied the spatial organization of the neural entities implied in different cognitive tasks, with a special interest for the visual system. The three main methodological contributions of this thesis are the following:

- A *Bayesian* approach for sparsity-inducing regularization, called *Multi-Class Sparse Bayesian Regression – MCBR* [47]. This approach is a generalization of the two principal Bayesian regularization techniques, *Bayesian Ridge Regression* and *Automatic Relevance Determination*.
- He proposed an approach, called *supervised clustering*, that includes spatial information in the prediction framework, and yields clustered weighted maps [25]. It can be used with any prediction functions for highly dimensional data.
- He implemented both sparsity and spatially-informed regularization within the same framework. He proposed a generalization of the *Total Variation regularization* for prediction task, and he showed its good performance in the case of *fMRI* data analysis [48].

In addition to these methodological directions, Vincent Michel has also worked on the implementation of the algorithms studied in detailed in this thesis and contributed to *Scikit-learn*, an open-source library for statistical learning (<http://scikit-learn.sourceforge.net/>).

Moreover, Vincent Michel and Robin Genuer examine the value of random forests to deal with such problems. [41] and present a new approach for the prediction of a behavioral variable from Functional Magnetic Resonance Imaging (fMRI) data. The difficulty comes from the huge number of image voxels that may provide relevant information with respect to the limited number of available images. Based on Random Forests, the approach provides an accurate auto-calibrated framework for selecting a reduced set of jointly informative regions.

6.6. Nonlinear mixed effects model

Participants: Marc Lavielle, Kevin Bleakley, Maud Delattre, Cyprien Mbogning, Benoît Charles.

Mixed effects hidden Markov models: MHMM have been recently developed as an extension of hidden Markov models to population studies. In the pharmacometric area, mixed hidden Markov models would provide an accurate description of longitudinal data collected during certain clinical trials, especially when distinct (hidden) disease stages are supposed to condition the distribution of some biological markers. Those particular models are quite easily interpretable and could even show similarities in the biological process that governs certain pathologies.

Our work mainly aimed at developing and evaluating a complete methodology for estimating parameters in those new models. Our algorithms were applied in the clinical context of epilepsy, to model daily seizure counts in epileptic patients and to assess the effects of a given anti-epileptic drug on the evolution of the epileptic symptoms.

Mixture models and model mixtures: a patient population is usually heterogeneous with respect to response to drug therapy. In any clinical efficacy trial, patients who respond, those who partially respond and those who do not respond present very different profiles. Then, diversity of the observed kinetics cannot be explained adequately only by the inter-patient variability of some parameters and mixtures are a relevant alternative in such situations.

We have extended the SAEM algorithm for mixture models and model mixtures. We have applied the proposed methodology for analyzing viral load data arising from HIV infected patients. We propose to describe these viral load data with a mixture of three models. Indeed, the data seem to exhibit three different typical profiles: responders, non-responders and rebounders.

Non-Linear Mixed Effects Models with Stochastic Differential Equations: the use of stochastic differential equations in non linear mixed effects models enables the decomposition of the intra-patient variability into some residual errors and some dynamical system variability.

We have developed a new algorithm based on the Stochastic Approximation EM (SAEM) method with the Kalman Filter for linear SDE systems and with the Extended Kalman Filter for nonlinear SDE systems.

7. Contracts and Grants with Industry

7.1. Contracts with EDF

Participants: Gilles Celeux, Jean-Michel Poggi.

- SELECT has a contrat with EDF regarding modelling uncertainty in deterministic models.
- SELECT has a contrat with EDF regarding wavelet analysis of the electrical load consumption for the aggregation and desaggregation of curves to improve total signal prediction.

7.2. The Monolix software project

Participant: Marc Lavielle.

Several pharmaceutical companies already joined the Monolix software project and signed a contract with INRIA:

- Novartis
- Roche
- Sanofi-Aventis
- Johnson & Johnson
- Astrazeneca
- Exprimio

7.3. Other contracts

Participants: Pierre Connault, Pascal Massart, Jean-Michel Poggi.

- SELECT has a contrat with SAFRAN - MESSIER-DOWTY, an high-technology group (Aerospace propulsion, Aircraft equipment, Defense Security, Communications), regarding modelling reliability of Aircraft Equipment (collaboration with Patrick Pamphile (Université Paris-Sud).
- SELECT has a contract with Total regarding short time Fourier transform for Spurious signal detection.

7.4. Project GAS

Participants: Gilles Celeux, Pascal Massart.

The project GAS was selected by the DIGITEO consortium in the framework of the “Domaines d’Intérêt Majeur” call of the Région Ile-de-France. The main partner is GEOMETRICA. The other partners of the project are the Ecole Polytechnique (F. Nielsen) and SELECT. The project intends to explore and to develop new researches at the crossing of information geometry, computational geometry and statistics. It started in September 2008 and its expected duration is two years. In this setting, Pascal Massart is the cosupervisor with Frédéric Chazal (GEOMETRICA) of the thesis of Claire Caillerie (GEOMETRICA).

7.5. Project EHPOC

Participants: Gilles Celeux, Rémi Fouchereau.

Select has been involved in the project EHPOC (industrial platform deliveries) of the pole SYSTEMATIC. An engineer, Rémi Fouchereau has been appointed to achieve a complete study on Bayesian inference for inverse problems in industrial uncertainty analysis.

8. Other Grants and Activities

8.1. National Actions

SELECT is animating a working group on model selection and statistical analysis of genomics data with the Biometrics group of Institut Agronomique Nationale Paris-Grignon (INAPG).

Pascal Massart is co-organizing a working group at ENS (Ulm) on Statistical Learning. This year the group focused interest on regularisation methods in regression. Most of SELECT members are involved in this working group.

SELECT is animating a working group on Classification, Statistics and fMRI imaging with Neurospin.

8.2. European actions

Gilles Celeux and Pascal Massart are members of the PASCAL (Pattern Analysis, Statistical Learning and Computational Learning) network.

8.3. International Initiatives

Gilles Celeux is one of the co-organizers of the Working Group on Model-Based Clustering.

9. Dissemination

9.1. Scientific Community animation

9.1.1. Editorial responsibilities

Participants: Gilles Celeux, Pascal Massart, Jean-Michel Poggi.

- Gilles Celeux is Editor-in-Chief of *Statistics and Computing*. He is Associate Editor of *CSBIGS* and *La Revue Modulad*.
- Pascal Massart is Associated Editor of *Annals of Statistics*, *Confluentes Mathematici*, and *Foundations and Trends in Machine Learning*.
- Jean-Michel Poggi is Associated Editor of *Journal of Statistical Software*, *Journal de la SFdS* and *CSBIGS*.

9.1.2. Invited conferences

Participants: Gilles Celeux, Pascal Massart, Marc Lavielle, Jean-Michel Poggi.

- Gilles Celeux was invited speaker to the Fall session of the Swiss Statistical Seminar in Bern and to the Summer Model-Based Clustering working group in Grenoble.
- Marc Lavielle was invited speaker at MAS 2010 in Bordeaux, Population PK 2010 in Amsterdam, participant of invited sessions at ISB 2010 in Montpellier, ERCIM 2010 in London.
- Jean-Michel Poggi was invited speaker at SIS 2010, Scientific Meeting of the Italian Statistical Society in Padua.

9.1.3. Scientific animation

Participants: Gilles Celeux, Pascal Massart, Marc Lavielle, Jean-Michel Poggi.

- Gilles Celeux is member of the scientific council of the MIA Department of INRA. He has been Guest Editor of the Special Issue on Mixture Models of the *Revue Modulad* (Number 40, July 2009). He was member of the AERES evaluation council of the Department BIAT (Unité de Biométrie et intelligence artificielle) of INRA Toulouse.
- Marc Lavielle is member of the Haut Conseil des Biotechnologies.
- Marc Lavielle is director of the GDR (Groupement de Recherche) "Statistique et Santé", Research Unit 3067 of the CNRS.
- Marc Lavielle is member of the council of the SMAI (Société de Mathématiques Appliquées et Industrielles).
- Marc Lavielle is member of the ITMO Committee on Public Health.
- Marc Lavielle was the head of the AERES evaluation committee of the MISTEA team,
- Marc Lavielle was member of the evaluation committee of the mathematical department of the Université Libre de Bruxelles.

- Pascal Massart is a member of the scientific council of the French Mathematical Society.
- Pascal Massart is a member of the scientific council of the Mathematical Department of the Ecole Normale Supérieure de Paris.
- Pascal Massart was a member of the scientific committee of the European Meeting of Statisticians in Piraeus.
- Jean-Michel Poggi is Cochair seminar of Probability and Statistics of the "laboratoire de Mathématiques d'Orsay", seminar ECAIS (Extraction de connaissances : approches informatiques et statistiques) of IUT de Paris 5 Descartes and of "Séminaire Parisien de Statistique".
- Jean-Michel Poggi is member of the Council of the French statistical society (SFdS).
- Jean-Michel Poggi is member of the Board of the "Environment group" of the French statistical society (SFdS).

9.2. Teaching

All the SELECT members are teaching in various courses of different universities and in particular in the M2 "Modélisation stochastique et statistique" of University Paris-Sud.

10. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

- [1] P. BARBILLON. *Méthodes d'interpolation à noyaux pour l'approximation de fonctions type boîte noire coûteuses*, Université Paris-sud 11, 2010.
- [2] R. GENUER. *Forêts aléatoires : aspects théoriques, sélection de variables et applications*, Université Paris-sud 11, 2010.
- [3] V. MICHEL. *Améliorer la compréhension du cortex visuel à l'aide de techniques de classification*, Université Paris-sud 11, 2010.
- [4] W. TOUSSILE. *Variable selection: Population genetic structure and Plasmodium transmission through mosquito*, Université Paris-sud 11, 2010.

Articles in International Peer-Reviewed Journal

- [5] M. AMINGHAFARI, J.-M. POGGI. *Multistep Forecasting Non-Stationary Time Series using Wavelets and Kernel Smoothing*, in "Communications in Statistics, Theory and Methods", 2010, to appear.
- [6] S. ARLOT, G. BLANCHARD, É. ROQUAIN. *Some non-asymptotic results on resampling in high dimension, I: confidence regions*, in "Annals of Statistics", 2010, vol. 38, p. 51-82.
- [7] S. ARLOT, G. BLANCHARD, É. ROQUAIN. *Some non-asymptotic results on resampling in high dimension, II: multiple tests*, in "Annals of Statistics", 2010, vol. 38, p. 83-99.
- [8] F. AUTIN, E. LE PENNEC, J.-M. LOUBES, V. RIVOIRARD. *Maxisets for Model Selection Estimators*, in "CA", 2010, vol. 31, n^o 2, p. 195–229.

- [9] F. AUTIN, E. LE PENNEC, K. TRIBOULEY. *Thresholding methods to estimate the copula density*, in "JMVA", 2010, vol. 101, n^o 1, p. 200–222.
- [10] P. BARBILLON, G. CELEUX, A. GRIMAUD, Y. LEFEBVRE, E. DE ROCQUIGNY. *Nonlinear methods for inverse statistical problems*, in "Computational Statistics & Data Analysis", 2011, vol. 55, n^o 1, p. 132–142.
- [11] J.-P. BAUDRY, A. RAFTERY, G. CELEUX, K. LO, R. GOTTARDO. *Combining mixture components for clustering*, in "Journal of Computational and Graphical Statistics", 2010, vol. 19, n^o 2, p. 332–353.
- [12] K. BERTIN, E. LE PENNEC, V. RIVOIRARD. *Adaptive Dantzig density estimation*, in "Annales de l'IHP", 2010, Accepted.
- [13] C. BIERNACKI, G. CELEUX, G. GOVAERT. *Exact and Monte Carlo calculations of integrated likelihoods for the latent class model*, in "JSPI", 2010, vol. 140, p. 2991–3002.
- [14] S. BOUCHERON, P. MASSART. *A high-dimensional Wilks phenomenon*, in "Probability Theory and Related Fields", 2010, p. 1–29.
- [15] P. CHAN, M. LAVIELLE, L. MCFADYEN, B. WEATHERLEY. *The Use of the SAEM Algorithm in MONOLIX Software for Estimation of Population Pharmacokinetic-Pharmacodynamic-Viral Dynamics Parameters of Maraviroc in Asymptomatic HIV Subjects*, in "Journal of Pharmacokinetics and Pharmacodynamics", 2010, to appear.
- [16] M. DELATTRE. *Inference in Mixed Hidden Markov Models and Applications to Medical Studies*, in "Journal de la SFdS", 2010, vol. 151, p. 90–105.
- [17] R. GENUER, J.-M. POGGI, C. TULEAU. *Variable Selection using Random Forests*, in "Pattern Recognition Letters", 2010, vol. 31, p. 2225–2236.
- [18] A. IACOBUCCI, J.-M. MARIN, C. ROBERT. *On variance stabilisation by double Rao-Blackwellisation*, in "Comput. Statist. Data Anal.", 2010, vol. 54, p. 698–710.
- [19] E. JOUINI, J.-M. MARIN, C. NAPP. *Discounting and Divergence of Opinion*, in "Journal of Economic Theory", Mar 2010, vol. 145, p. 830–859, <http://hal.inria.fr/halshs-00176636>.
- [20] G. KERKYACHARIAN, G. KYRIAZIS, E. LE PENNEC, P. PETRUSHEV, D. PICARD. *Inversion of noisy Radon transform by SVD based needlets*, in "ACHA", 2010, vol. 28, n^o 1, p. 24–45.
- [21] G. KERKYACHARIAN, E. LE PENNEC, D. PICARD. *Radon needlet thresholding*, in "Bernoulli", 2010, Accepted.
- [22] M. LAVIELLE, A. SAMSON, A. KARINA FERMIN, F. MENTRÉ. *Maximum Likelihood Estimation of Long-Term HIV Dynamic Models and Antiviral Response.*, in "Biometrics", May 2010, <http://hal.inria.fr/inserm-00486937>.
- [23] P. MASSART, C. MEYNET. *An ℓ_1 -oracle inequality for the Lasso*, in "Electronic Journal of Statistics", 2010, to appear, <http://hal.archives-ouvertes.fr/docs/00/50/64/46/PDF/RR-7356.pdf>.

- [24] C. MAUGIS, B. MICHEL. *A non asymptotic penalized criterion for Gaussian mixture model selection*, in "ESAIM, P & S", 2011, vol. 15.
- [25] V. MICHEL, A. GRAMFORT, G. VAROQUAUX, E. EGER, C. KERIBIN, B. THIRION. *A supervised clustering approach for fMRI-based inference of brain states*, in "Pattern Recognition - Special Issue on Brain Decoding", 2010, (to appear).
- [26] M. MISITI, Y. MISITI, G. OPPENHEIM, J.-M. POGGI. *Optimized Clusters for Disaggregated Electricity Load Forecasting*, in "REVSTAT", 2010, vol. 8, p. 105-124.
- [27] R. SAVIC, F. MENTRE, M. LAVIELLE. *Implementation and Evaluation of an SAEM algorithm for longitudinal ordered categorical data with an illustration in pharmacometrics*, in "The AAPS Journal", 2010, to appear.
- [28] E. SNOECK, P. CHAN, M. LAVIELLE, P. JACQMIN, N. JONSSON, K. JORGA, T. GOGGIN, S. JUMBE, N. FREY. *Hepatitis C Viral Dynamics Explaining Breakthrough, Relapse or Response after Chronic Treatment*, in "Clinical Pharmacology and Therapeutics (AAPS Outstanding Manuscript Award in Modeling and Simulation)", 2010, vol. 87, p. 706-713.
- [29] N. VERZELEN. *Adaptive estimation of Stationary Gaussian fields*, in "Annals of Statistics", 2010, vol. 38, p. 1363-1402.
- [30] N. VERZELEN. *High-dimensional Gaussian model selection on a Gaussian design*, in "Ann. Inst. H. Poincaré Probab. Statist.", 2010, vol. 46, p. 480-524.
- [31] N. VERZELEN, F. VILLERS. *Goodness-of-fit Tests for high-dimensional Gaussian linear models*, in "Annals of Statistics", 2010, vol. 38, p. 704-752.

Articles in National Peer-Reviewed Journal

- [32] C. KERIBIN. *Méthodes bayésiennes variationnelles : concepts et applications en neuroimagerie*, in "Journal de la Société Française de Statistiques", 2010, (à paraître).

Invited Conferences

- [33] A. ANTONIADIS, X. BROSSAT, J. CUGLIARI, J.-M. POGGI. *Functional Clustering using Wavelets*, in "Proceedings of the 19th COMPSTAT", Paris, August 2010, p. 697-704.
- [34] G. CELEUX. *Comparing different points of view for analyzing finite mixture models*, in "Swiss Statistics Seminar Fall Session", Bern, November 2010.
- [35] G. CELEUX. *Comparing variational and stochastic algorithms for the maximum likelihood estimation of model-based block clustering*, in "Working Group on Model-Based Clustering Summer Session", Grenoble, July 2010.
- [36] J. CUGLIARI. *Forecasting power electricity demand using a functional non parametric model*, in "Workshop Industry & Price Forecasting, EDF R&D", Clamart, June 2010.
- [37] F. MHAMDI, M. JAIDANE, J.-M. POGGI. *Empirical Mode Decomposition for Trend Extraction. Application to Electrical Data*, in "Proceedings of the 19th COMPSTAT", Paris, August 2010, p. 1391-1398.

International Peer-Reviewed Conference/Proceedings

- [38] A. ANTONIADIS, X. BROSSAT, J. CUGLIARI, J.-M. POGGI. *Classification non supervisée des données fonctionnelles avec ondelettes*, in "42èmes Journées de Statistique", France Marseille, France, 2010, <http://hal.inria.fr/inria-00494715>.
- [39] P. BARBILLON, G. CELEUX, A. GRIMAUD, Y. LEFEBVRE, E. DE ROCQUIGNY. *Méthodes non linéaires pour des problèmes statistiques inverses*, in "42èmes Journées de Statistique", France Marseille, France, 2010, <http://hal.inria.fr/inria-00494690>.
- [40] R. GENUER. *Bornes de risque pour les forêts purement uniformément aléatoires*, in "42èmes Journées de Statistique", France Marseille, France, 2010, <http://hal.inria.fr/inria-00494837>.
- [41] R. GENUER, V. MICHEL, E. EGER, B. THIRION. *Random Forests based feature selection for decoding fMRI data*, in "Proceedings of the 19th COMPSTAT", November 2010, p. 1071-1078.
- [42] C. KERIBIN, G. GOVAERT, G. CELEUX. *Estimation d'un modèle à blocs latents par l'algorithme SEM*, in "42èmes Journées de Statistique", France Marseille, France, 2010, <http://hal.inria.fr/inria-00494796>.
- [43] C. KERIBIN, G. GOVAERT, G. CELEUX. *Estimation d'un modèle à blocs latents par l'algorithme SEM*, in "42èmes Journées de Statistique", Marseille, France, 2010.
- [44] V. MICHEL, E. EGER, C. KERIBIN, J.-B. POLINE, B. THIRION. *A supervised clustering approach for extracting predictive information from brain activation images*, in "Workshop on Mathematical Methods in Biomedical Image Analysis - IEEE Conference on Computer Vision and Pattern Recognition", États-Unis San Francisco, Jun 2010, 08, <http://hal.inria.fr/hal-00504094>.
- [45] V. MICHEL, E. EGER, C. KERIBIN, B. THIRION. *Multi-Class Sparse Bayesian Regression for Neuroimaging data analysis*, in "International Workshop on Machine Learning in Medical Imaging (MLMI) In conjunction with MICCAI 2010", Chine Beijing, Sep 2010, 1, <http://hal.inria.fr/hal-00505057>.
- [46] V. MICHEL, E. EGER, C. KERIBIN, J.-B. POLINE, B. THIRION. *A supervised clustering approach for extracting predictive information from brain activation images*, in "IEEE Computer Society Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA10) Workshop on Mathematical Methods in Biomedical Image Analysis - IEEE Conference on Computer Vision and Pattern Recognition", 06 2010, 08.
- [47] V. MICHEL, E. EGER, C. KERIBIN, B. THIRION. *Multi-Class Sparse Bayesian Regression for Neuroimaging data analysis*, in "International Workshop on Machine Learning in Medical Imaging (MLMI) 2010 In conjunction with MICCAI 2010", 09 2010, p. 50-57.
- [48] V. MICHEL, A. GRAMFORT, G. VAROQUAUX, B. THIRION. *Total Variation regularization enhances regression-based brain activity prediction*, in "1st ICPR Workshop on Brain Decoding - Pattern recognition challenges in neuroimaging - 20th International Conference on Pattern Recognition", 08 2010, 1.
- [49] A. PASANISI, S. FU, N. BOUSQUET. *Estimation de modèles markoviens discrets dans un cadre industriel fiabiliste à données manquantes.*, in "42èmes Journées de Statistique", 2010.

Workshops without Proceedings

- [50] P. BARBILLON, S. FU, R. FOUCHEREAU. *Problème inverse statistique dans des modèles de crues. Traitement par un algorithme EM et par une méthodologie bayésienne.*, in "Journée modèles inverses, SFdS", Paris, France, 2010.
- [51] S. FU, G. CELEUX, M. COUPLET, N. BOUSQUET. *Inversion probabiliste en analyse d'incertitude.*, in "Séminaire d'animation scientifique et technique de la compétence APSPP", 2010.
- [52] S. FU, G. CELEUX, M. COUPLET, N. BOUSQUET. *Inversion probabiliste en analyse d'incertitude.*, in "Séminaire des thèses, EDF R&D", 2010.

Research Reports

- [53] Y. AUFRAY, P. BARBILLON, J.-M. MARIN. *Maximin Design on non-hypercube domain and Kernel Interpolation*, arXiv, 2010, <http://arxiv.org/abs/1004.0784>.
- [54] P. BARBILLON, G. CELEUX, R. FOUCHEREAU, S. FU. *Nonlinear inference in Inverse Problems*, EHPOC Report (60 pages), 2010.
- [55] G. CELEUX, M. EL-ANBARI, J.-M. MARIN, C. P. ROBERT. *Regularization in regression: comparing Bayesian and frequentist methods in a poorly informative situation*, INRIA, Oct 2010, RR-7415, <http://hal.inria.fr/inria-00525436>.
- [56] R. GENUER. *Risk bounds for purely uniformly random forests*, INRIA, Jun 2010, RR-7318, <http://hal.inria.fr/inria-00492231>.
- [57] R. GENUER. *Risk bounds for purely uniformly random forests*, inria, 2010, <http://hal.archives-ouvertes.fr/inria-00492231/en/>.
- [58] V. VANDEWALLE, C. BIERNACKI, G. CELEUX, G. GOVAERT. *A predictive deviance criterion for selecting a generative model in semi-supervised classification*, INRIA, Sep 2010, RR-7377, <http://hal.inria.fr/inria-00516991>.

Other Publications

- [59] J.-P. BAUDRY. *Sélection de modèle pour la classification non supervisée*, Aug 2010, Type : Conference digest, <http://hal.inria.fr/inria-00510149>.
- [60] J.-P. BAUDRY, C. MAUGIS, B. MICHEL. *Pratique de l'heuristique de pente et le package CAPUSHE*, Aug 2010, Type : Conference digest, <http://hal.inria.fr/inria-00510367>.
- [61] J.-P. BAUDRY, C. MAUGIS, B. MICHEL. *Slope Heuristics: Overview and Implementation*, Mar 2010, <http://hal.inria.fr/hal-00461639>.
- [62] K. BERTIN, E. LE PENNEC, V. RIVOIRARD. *Estimation de densité par pénalisation ℓ_1 minimale*, 11 2010, GdT ISIS.

-
- [63] K. BERTIN, E. LE PENNEC, V. RIVOIRARD. *Estimation de densité par pénalisation ℓ_1 minimale*, 12 2010, 50 ans du LPMA.
- [64] E. LE PENNEC, S. COHEN. *Estimation de densités conditionnelles par sélection de modèle et application à la segmentation d'images hyperspectrales*, 11 2010, Séminaire Parisien de Statistique.
- [65] E. LE PENNEC, S. COHEN. *Segmentation d'images hyperspectrales*, 02 2010, GTOX.
- [66] E. LE PENNEC, S. COHEN. *Segmentation d'images hyperspectrales par mélange de gaussiennes et sélection de modèles*, 03 2010, ENSAE.
- [67] E. LE PENNEC, S. COHEN. *Segmentation d'images hyperspectrales par mélange de gaussiennes et sélection de modèles*, 09 2010, Orsay.
- [68] E. LE PENNEC, S. COHEN. *Segmentation d'images hyperspectrales par mélange de gaussiennes et sélection de modèles*, 10 2010, Geometrica.
- [69] E. LE PENNEC, S. COHEN. *Segmentation d'images hyperspectrales par mélange de gaussiennes et sélection de modèles*, 10 2010, GTOX.
- [70] E. LE PENNEC, J. SALMON. *An aggregated point of view on NL-Means*, 5 2010, Modélisation Statistique des Images.
- [71] C. MAUGIS, G. CELEUX, M.-L. MARTIN-MAGNIETTE. *Variable selection in model-based discriminant analysis*, RR-7290, <http://hal.inria.fr/inria-00483229>.