INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team select

Model Selection and Statistical Learning

Saclay - Île-de-France

Theme : Optimization, Learning and Statistical Methods

Table of contents

1. Team

Research Scientist

Gilles Celeux [Team Vice-Leader, DR INRIA]
Marc Lavielle [DR INRIA detached from Université Paris 5]
Erwan Le Pennec [CR INRIA since October 2009]

Faculty Member

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

External Collaborator

Yves Auffray [Dassault]

Technical Staff

Benoît Charles
Kaelig Chatel
Morgan Guery
Hector Mesa
Jean-François Si Abdallah

PhD Student

Pierre Barbillon [MESR grant]
Jean-Patrick Baudry [MESR grant]
Pierre Connault [CIFRE grant]
Jairo Cugliari-Duhalde [CIFRE grant]
Maud Delattre [Région IDF grant]
Mohammed El Anbari [France-Marocco grant]
Robin Genuer [MESR grant]
Merlin Keller [CEA-INRIA grant]
Julie Marcelin [ENS grant]
Cyprien Mbogning [INRIA grant]
Bertrand Michel [CIFRE grant]
Vincent Michel [INRIA grant]
Wilson Toussile [France-Cameroun grant]
Vincent Vandewalle [MESR grant]

Post-Doctoral Fellow

Cathy Maugis [ATER]
Nicolas Verzelen [ENS grant]

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.

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. Beyond the specification of the joint distribution, the Bayesian approach is automatic. 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. The SELECT team is interested in applications of this Bayesian approach for model uncertainty problems where a large number of different models are under consideration. The joint distribution is obtained by introducing prior distributions on all the unknowns, here the parameters of each model and the models themselves, and then combining them with the distributions for the data. Conditioning on the data then induces a posterior distribution of model uncertainty that can be used for model selection and other inference and decision problems. This is the essential idea and it can be powerful. However, two major challenges confront its practical implementation: the specification of the prior distributions and the calculation of various posterior distributions.

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, neuroimaging and phylogeny.

4.2. Curves classification

The field of classification for complex data as curves, functions, spectra and time series is important in situations when the values of the explanatory variables of each value are functional, rather than scalar. Classic data analysis questions are being revisited to define new strategies that take the functional nature of the data into account. This new domain, functional data analysis, addresses a variety of applied problems, including longitudinal studies, analysis of fMRI data and spectral calibration.

We are focusing on classification problems with a particular emphasis on clustering, i.e. unsupervised classification. In addition to classic questions such 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 requires a very large computational effort, which need to be addressed with efficient or anytime algorithms.

4.3. Reliability

Since several years, SELECT has collaborations with EDF-DER *Fiabilité des Composants et Structures* group. An important theme was the problem of aging modelling of nuclear material components in order to analyse the durability of nuclear plants. Since 2007 a new theme has been introduced in this collaboration. It concerns the resolution of inverses problems using simulation tools to analyse incertainty in highly complex physical systems.

The other major theme concerns fatigue rupture analysis based on a research collaboration based on a research collaboration with SAFRAN an high-technology group (Aerospace propulsion, Aicraft equipment, Defense Security, Communications). The aim is to perform an efficient statistical control of aircraft equipment production processes.

The LASSO is a selection method for linear regression. It minimizes the sum of squared errors, with a penalty on the sum of the absolute values of the coefficients. The objectif of Pierre Connault, in his PhD, is to calibrate automatically that penalty.

The fatigue strength of aeronotics equipments is tested from cyclic applications of stress on small steel test-tubes. How do these results on test-tubes can be extrapolated to reliability analysis of the entire equipment ? A probabilistic model is proposed by SELECT.

4.4. Phylogeny

Phylogeny is concerned with designing evolutionary trees between species from aligned nucleotide sequences. More precisely, a nucleotide sequence being an ordered set of sites taking value in a finite set E (for instance, $E = \{A, C, G, T\}$), the problem is to reconstruct the topology of the evolutionary tree between the species from aligned sequences for the considered species, and to estimate the tree parameters (branches length) as well as the parameters of the evolutionary model. Our research in this domain is twofold. First we are working on a model selection approach from a semi parametric graphical model whose parameters to be estimated are the topology, branches lengths and mutation rate of the evolutionary tree. Secondly, we are working on the *covarion* model. For this model, a site can change behavior along the evolutionary tree according to two hidden states, active (ON) or nonactive (OFF). In this research, we are interested in comparing non nested models.

4.5. Population genetics

SELECT develops new methods of statistical inference on molecular data obtained from population samples. Some of these methods are aimed at treating complex evolutionary scenarios, including several populations related by phylogenetic trees, with possible admixture and/or migration. Other methods will explicitly take into account the spatial distribution of samples. Inference concerns the parameters of these scenarii, which mainly characterize the population demographic history and the mutation model of markers. The explicit use of geographic information allows for a more efficient characterization of evolutionary episodes poorly analyzed by existing methods, such as bioinvasions or shifts of species distribution areas due to global climatic changes. The analysis of complex scenarii combines two algorithms: an Importance Sampling algorithm to estimate the data likelihood under a given scenario and with given values of parameters and a second algorithm (to be determined) to explore efficiently the parameter space.

4.6. 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.

4.7. 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.

4.8. Computer Experiments

Since 2007, SELECT developed several computer experiment studies, in the framework of conventions with Dassault Aviation and EDF. They concern the resolution of inverses problems using simulation tools to analyse incertainty in highly complex physical systems.

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 peerformances, code a proper graphical library for clustering displays in MIXMOD and propose a version available via internet. On the other MIXMOD received the support of Digiteo to investigate the possibility of a commercial diffusion of this software aside its free diffusion.

5.2. MONOLIX software

Participants: Marc Lavielle [Correspondant], Benoît Charles, Kaelig Chatel, Morgan Guery, 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, Saint-Petersbourg, June 2009,
- ACOP meeting, Mystic (US), October 2009,
- Buffalo University (US), March 2009,

Version 3.1 of MONOLIX is available since October 2009. This version of the software was developed thanks to the financial support of Novartis, Roche, J&J, Sanofi-Aventis, Exprimo. This version was presented during the MONOLIX Day on November 16th at the Maison de la Recherche, Paris.

The MONOLIX Project consists primarily in 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 Développment Logiciel) to hire two engineers (Kaelig Chatel until August 2009, Hector Mesa).

We have obtained from DIGITEO an OMTE (Opération de Maturation Technico-Economique) to hire one engineer (Kaelig Chatel from September 2009), 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, Pascal Massart, Cathy Maugis, Bertrand Michel, Jean-Michel Poggi, Vincent Vandewalle, Nicolas Verzelen.

In collaboration with Marie-Laure Martin-Magniette (URGV et UMR AgroParisTech/INRA MIA 518), Gilles Celeux and Cathy Maugis developed a variable selection procedure for model-based clustering in [18]. The problem is regarded as a model selection problem in the model-based cluster analysis context. They proposed a more versatile variable selection model taking into account three possible roles for each variable: The relevant clustering variables, the irrelevant clustering variables dependent on a part of the relevant clustering variables and the irrelevant clustering variables totally independent of all the relevant variables. This modelling allows to generalize the model of [17] and the one of Raftery and Dean (2006). A model selection criterion based on BIC and a variable selection algorithm called *SelvarClustIndep*, embedding two backward stepwise variable selection algorithms for clustering and linear regression, are derived for this new variable role modelling. The model identifiability and the consistency of the variable selection criterion are also established. Numerical experiments highlight the interest of this new modeling. Two softwares, so-called *SelvarClust* and *SelvarClustIndep*, implemented en C^{++} are devoted to the variable selection in model-based clustering according to the modelling of [17] and [18] respectively. They are available at the following address: http://www.math.univ-toulouse.fr/~maugis. Currently, those researchers are interested in taking advantage of this general variable role modelling for discriminant analysis. In particular, the variable selection gives a new interest for the quadratic discriminant analysis.

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).

Cathy Maugis with Bertrand Michel (GEOMETRICA, Inria) consider specific Gaussian mixtures to solve simultaneously variable selection and clustering problems. In [19], they proposed a non asymptotic penalized criterion to choose the number of mixture components and the relevant variable subset. Because of the non linearity of the associated Kullback-Leibler contrast on Gaussian mixtures, a general model selection theorem for MLE proposed by Massart is used to obtain the penalty function form and the associated oracle inequality. This theorem requires controlling the bracketing entropy of mixture families. Nevertheless, these theoretical results depend on unknown constants. Currently, they are interested in establishing an adaptative property of their penalized maximum likelihood estimators in a minimax sense. In [67], they study the practical use of their penalized criterion. A "slope heuristics" method is applied to calibrate these constants. Jean-Patrick Baudry, Cathy Maugis and Bertrand Michel [49] are developping 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 ready-to-use and easy solution for people who may want to try to apply the slope heuristics and then to encourage its use. They are preparing an overview about the slope heuristics to introduce this package.

With Sylvain Arlot (ENS, CNRS), Pascal Massart [7] studied the so-called slope heuristics in the framework of regression on a random design, with possible heteroscedastic noise. Assuming that all the models are made of histograms, they show the same relationship between a “minimal penalty” and an optimal one. This can for instance be used for tuning a penalty, when the optimal penalty is known up to some multiplicative constant. In general, the optimal shape of the penalty can be estimated by v-fold or resampling penalties. Their work

is based on new structural concentration inequalities for the empirical risk and the high-dimensional Wilks phenomenon enlightened in [9].

Jean-Patrick Baudry and Gilles Celeux [1] continued the study of estimation and model selection procedures derived by minimizing a new contrast adapted for clustering with mixture models and inspired from the ICL criterion. Theoretical results have been obtained about the consistency of the estimator thus defined and about the consistency of the corresponding model selection procedure. Moreover, solutions have been developped to practically compute this estimator, which involves difficulties analagous to those arising when computing the usual maximum likelihood estimator with mixture models. Those solutions may improve the results of the EM algorithm in this usual task, too. They also studied some robustness properties of the proposed estimator.

Jean-Patrick Baudry and Gilles Celeux, in collaboration with Ana Maria Ferreira (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*. This criterion has been applied to a data set in the professional development field.

In collaboration with Professor Abdallah Mkhadri (University of Marrakesh, Marocco), Gilles Celeux supervised the thesis of Mohammed El Anbari which concern regularisation methods in linear regression. This year, in collaborarion with Jean-Michel Marin (Université de Montpellier) they have considered the Bayesian point of view and compared Bayesian methods of variable seletion in linear regression with standard regularisation methods in a poorly informative context [53].

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 is articulated following three directions:

- The preliminary theoretical direction concerns mathematical understanding of the reasons of this amazing behaviour.
- The second methodological direction aims at improving the knowledge about how to tune the parameters. It includes computer intensive simulations and comparisons based on well-known real data sets.
- 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.

In [41], Robin Genuer and Vincent Michel 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.

6.2. Selection of high dimensional graphical models

Participants: Pascal Massart, Nicolas Verzelen.

The last decade has witnessed the apparition of applied problems typified by very high-dimensional variables (in marketing database or gene expression studies for instance). Graphical models enable concise representations of associational relations between variables. If the graph is known, the parameters of the model are easily estimated. However, a quite challenging issue is the selection of the most appropriate graph for a given data set.

Sylvie Huet (INRA), Pascal Massart, Nicolas Verzelen, and Fanny Villers (Université Paris 6) [24] defined a goodness-of-fit test of linear hypotheses for Gaussian regression with Gaussian covariates. They deduced from it a test for Gaussian graphical models which applies in a high dimensional setting. Besides, it is shown to be minimax against various alternatives. They have also carried out numerical experiments with microarray genetic data and have assessed the graph of genetic networks [25].

Graph selection of Gaussian graphical models is closely related to the estimation in the linear regression model with Gaussian covariates. In this setting, Nicolas Verzelen [23] has introduced a novel estimation method based on penalization ideas. This procedure is proved to satisfy a non-asymptotic oracle inequality and adaptation properties. Contrary to other methods such as the lasso, the rates of convergence do not depend on the correlation between the covariates.

Verzelen's procedure [23] allows to tackle graph selection. However, its computational cost becomes prohibitive when the size of the graph increases. To handle this drawback, Christophe Giraud (École Polytechnique), Sylvie Huet (INRA), and Nicolas Verzelen propose a two-stage procedure which first builds a family of candidate graphs from the data and then selects one graph among this family according to a dedicated criterion [65]. This estimation procedure is shown to be consistent in a high-dimensional setting and its risk is controlled by a non-asymptotic oracle-like inequality. A nice behavior on numerical experiments corroborates these theoretical results. The procedure is implemented in the R-package GGMselect available on the CRAN.

6.3. Statistical learning methodology and theory

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

Gilles Celeux and Jean-Patrick Baudry, in collaboration with Adrian Rafetry, Kenneth Lo and Raphael Gottardo [64], proposed a methodology for clustering which is an attempt to take advantage of mixture models both for their usefulness in clustering and for their good approximation properties — by modeling each class itself as a mixture. The question which has to be answered is then how to choose the mixture components to gather. They proposed to consider a criterion based on an entropy measure of the obtained classification quality. The resulting hierachical methodology is notably illustrated by its application to a flow cytometry data set.

In collaboration with Christophe Biernacki (Université de Lille) and Gérard Govaert (UTC Compiègne), Gilles Celeux propose non asymptotic version of integrated likelihoods for the latent class model or multivariate multinomial mixture model. They exploit the fact that a fully Bayesian analysis with Jeffreys non informative prior distributions does not involve technical difficulty to propose an exact expression of the integrated *complete-data* likelihood, which is known as being a meaningful model selection criterion in a clustering perspective. Moreover, this year they propose importance sampling strategies taking into acount the so-called label switching problem to get efficient approximations of the integrated *observed-data* likelihood.

Vincent Vandewalle will defend his PhD these in december 2009 about semi-supervised model-based classification under the supervision of Christophe Biernacki (Université de Lille), Gilles Celeux and Gérard Govaert(UTC). His thesis focused on the discriminant analysis situation which is of main interest for applications. Firstly, he designed an hypothesis test to take profit of unlabeled data to decide if a classification model is reliable. Then, he has conceived and investigated specific information based criteria for model selection in the semi-supervised setting. Conceived in the same spirit than the BEC criterion of Bouchard and Celeux (2006)[1] by taking into account the classification purpose, he proposed an AIC-like criterion which behaves slighly better in practice and has better theoretical features. In the supervised setting, he proposed an alternative AIC-like criterion which penalises the conditionnal data likelihood by the number of independent parameters involved in the conditionnal likelihood when a generative model is learned [45]. This criterion has shown a promising behavior and is cheap.

[1] IEEE on PAMI

Jean-Michel Poggi proposed a procedure for detecting outliers in regression problems. It is based on information provided by boosting regression trees. The key idea is to select the most frequently resampled observation along the boosting iterations and reiterate after removing it. The selection criterion is based on Tchebychev's inequality applied to the maximum over the boosting iterations of the average number of appearances in bootstrap samples. Thus, the procedure is noise distribution free. A lot of well-known bench data sets are considered and a comparative study against two well-known competitors allows to show the interest of the method.

6.4. Reliability and Computer Experiments

Participants: Yves Auffray, Pierre Barbillon, Gilles Celeux, Pierre Connault, Pascal Massart.

In the framework of a convention with EDF, Gilles Celeux worked in collaboration with Yannick Lefebvre and Étienne de Rocquigny (EDF) on the resolution of not linear inverse problems for the quantification of uncertainties in a physical model. More precisely, noisy observed data (Y) were dependent, through a known but complex and expensive function H from non-observed data X. The aim is to estimate parameters of the probability distribution of the non observed data (X) and the variance of the noise. The problem has a missing data structure and can be solved with an EM-type algorithm coupled to an iterative linearisation of the function H ([10]). However, the linearisations of the function H can be poor and this scheme misleading. Pierre Barbillon, Gilles Celeux and Agnès Grimaud (Université de Marseille) proposed a non-linearised method coupling the use of the Stochastic EM algorithm with a MCMC method and a Kriging approximation of the H function. The algorithm and its practical figures are described in [63] and is compared with iterated linear approximation on the basis of numerical experiments on simulated and real data sets. Situations where this non linear approach is to be preferred to linearisation are highlighted.

In aircraft equipment, fatigue is one of the first cause of ruptures. Morevover fatigue ruptures appear brutally and can be catastrophics: important material damage, human death. The fatigue rupture is a complex random process: it is influenced by numerous and various factors of the production process and environment. The large number of factors, and the strong variability of some of them, yield any expertise very difficult. SELECT develops a collaboration with SAFRAN via the Phd of Pierre Connault, supervised by Pascal Massart and Patrick Pamphile (Université Paris-Sud). Variable selection methods (CART, LASSO) have been used to perform an efficient statistical control of aircraft equipment production processes. LASSO is a regularisation method for linear regression. It minimizes the sum of squared errors, with a penalty on the sum of the absolute values of the coefficients. The objectif of Pierre Connault is to calibrate automatically that penalty. Moreover, a probabilistic model of fatigue has been proposed to extrapolate results on tets tubes to assess the reliability of the entire equipment.

Yves Auffray and Pierre Barbillon have proposed a more natural and general definition of a conditionally positive definite kernel in [61]. From this definition, a full generalization of Aronszajna's theorem has been shown. It states for a conditionally positive definite kernel, the existence of a unique reproducing kernel semi-Hilbert space. Furthermore, they provided the interpolation operator on this space and showed that it is a generalization of the standard ones.

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. Yves Auffray and Pierre Barbillon have justified with Jean-Michel Marin (Université de Montpellier) to take a design satisfying to the MAXIMIN criterion by using results from the approximation theory literature. 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. It can be done thanks to a well-known algorithm of Morris and Mitchell (1992). However, the Latin hypercube sampling is pointless in a non hypercubic domain. In [62], they proposed a simulated annealing algorithm, implemented in *C*, which aims at obtaining a MAXIMIN design in any bounded connected domain. They have proved the convergence of their algorithm.

6.5. Classification in genomics

Participants: Gilles Celeux, Cathy Maugis.

Following the Cathy Maugis thesis in 2008, we decide to use her material in collaboration with biologists of URGV (INRA, Evry Genopole) and Marie-Laure Martin-Magniette (INRA) to improve functional annotation of *Arabidopsis thaliana* genes. This joint work with URGV is entering in the SONATA project which will be pussued in 2010.

This year, the variable selection procedures concieved by Cathy Maugis are in particular used for genomics applications which is the result of a collaboration with researchers of of URGV (Evry Genopole). Biologists are interested in predicting the gene functions of sequenced genome organisms according to microarray transcriptome data. The microarray technology development allows one to study the whole genome in different experimental conditions. The information abundance may seem to be an advantage for the gene clustering. However, the structure of interest can often be contained in a subset of the available variables. In [17], the variable selection algorithm *SelvarClust* was used to extract groups of coexpressed *Arabidopsis thaliana* genes. It allowed to improve the clustering and make easier the biological interpretation. In [26], the interest of the new variable selection algorithm *SelvarClustIndep* for discovering the function of orphan genes is highlighted on a transcriptome dataset for the *Arabidopsis thaliana* plant.

6.6. Curves classification, denoising and forecasting

Participant: Jean-Michel Poggi.

In collaboration with Anestis Antoniadis (Université J. Fourier, Grenoble) and Irène Gijbels (Leuven Unversity), Jean-Michel Poggi considered a non parametric noisy data model $Y_k = f(x_k) + \epsilon_k, \quad k = 1, \cdots, n,$ where the unknown signal f from $[0, 1]$ in $\mathbf{R}$ is assumed to belong to a wide range of function classes, including discontinuous functions and the $\epsilon_k' s$ are independent identically distributed noises with zero median. The unknown distribution of the noise is assumed to have heavy tails, so that no moments of the noise exist. The design points are assumed to be deterministic points, not necessarily equispaced within the interval $[0, 1]$. Standard kernel methods cannot be applied in this situation. Their approach first uses local medians to construct variables Z_k structured as a Gaussian nonparametric regression, then they apply a wavelet block penalizing procedure adapted to non equidistant designs to construct an estimator of the regression function. Under mild assumptions on the design, they show that their estimator, which has a good practical behavior, simultaneously attains the optimal rate of convergence over a wide range of Besov classes, without prior knowledge of the smoothness of the underlying functions or prior knowledge of the error distribution [3].

In order to take into account the variation of EDF (the French electrical company) portfolio due to the liberalization of the electrical market, it is essential to conveniently disaggregate the global signal. The idea is to disaggregate the global load curve in such a way that the sum of disaggregated predictions improve significantly the prediction of the global signal considered as a whole. In collaboration with Michel Misiti (Ecole Centrale de Lyon), Yves Misiti (Université Paris-Sud), G. Oppenheim (Université Marne la Vallée), Jean-Michel Poggi designs a strategy to optimize with respect to a predictability index, a preliminary clustering of individual load curves. The optimized clustering scheme is directed by forecasting performance via a cross-prediction dissimilarity index and proceeds as a discrete gradient type algorithm [68].

- Forecasting time series using wavelets :

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 comsumption 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.

6.7. Neuroimaging, Statistical analysis of fMRI data

Participants: Gilles Celeux, Robin Genuer, Merlin Keller, Christine Keribin, Marc Lavielle, 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 and Bertrand Thirion (Parietal). During his second year of thesis, he studied different ways to introduce spatial information in the features selection techniques, such as developing clustering technique using supervised information. Moreover, he developed an adaptive regularization method, which is estimated within a variational bayes framework ([44]). Bertrand Thirion and Vincent Michel have also been implied in different neuroscientific studies using classification techniques to improve the analysis of the data : relationship between regions of the brain during mental arithmetic ([15]), mental representation of quantities ([12]) and study of the valuation system in the human brain ([16]).

Moreover, Vincent Michel and Robin Genuer examine the value of random forests to deal with such problems.

Christine Keribin achieved a bibliographical study on the variational methods : describing the principle of variational methods and their applications in the Bayesian inference, surveying the main theoretical results and detailing two examples in the neuroimage field [66].

Merlin Keller began his PhD in October 2006 under the supervision of Alexis Roche (CEA, Neurospin) and Marc Lavielle. This thesis is dedicated to the statistical analysis of multi-subject fMRI data, with the purpose of identifying bain structures involved in certain cognitive or sensori-motor tasks, in a reproducible way across subjects. To overcome certain limitations of standard voxel-based testing methods, as implemented in the Statistical Parametric Mapping (SPM) software, a Bayesian model selection approach to this problem is used, meaning that the most probable model of cerebral activity given the data is selected from a pre-defined collection of possible models. Based on a parcellation of the brain volume into functionally homogeneous regions, each model corresponds to a partition of the regions into those involved in the task under study and those inactive. This allows to incorporate prior information, and avoids the dependence of the SPM-like approach on an arbitrary threshold, called the clusterforming threshold, to define active regions. By controlling a Bayesian risk, the approach balances false positive and false negative risk control. Furthermore, it is based on a generative model that accounts for the spatial uncertainty on the localization of individual effects, due to spatial normalization errors [42].

6.8. Nonlinear mixed effects model

Participants: Marc Lavielle, Maud Delattre, Julie Marcelin, Benoît Charles.

The MONOLIX group (http://software.monolix.org), co-chaired by Marc Lavielle, develops activities in the field of mixed effect models. This group involves scientists with different backgrounds, interested both in the study and applications of these models.

A promising collaboration started in 2009 with Pierre Del Moral (INRIA, ALEA) who is a well-known specialist in the field of stochastic algorithms. Collaborations with the MONOLIX team are mainly related with the estimation in mixed models defined by stochastic differential equations. For linear SDE systems, the Kalman filter is combined with the SAEM algorithm. Particle filters is a promising solution for non linear systems. Pierre Del Moral cosupervises with Marc Lavielle the PhD thesis of Julie Marcelin who is working on this topic of main interest for practical applications (PKPD models, glucose/insulin models for example).

This collaboration with the ALEA team is concerned with several other difficult (and important) problems:

- estimation in mixed hidden Markov models,
- simulated annealing for maximum likelihood estimation in mixed effects models,
- convergence of the SAEM algorithm in a general framework.

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
- Exprimo

Furthermore, Marc Lavielle collaborates with The MathWorks for implementing the SAEM algorithm in the Statistical Toolbox.

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, Aicraft 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 it is expected duration is two years. In this setting, Pascal Massart is the cosupervisor with Frédéric Chazal (GEOMETRICA) of the thesis of Claire Caillerie (GEOMETRICA).

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.

Jean-Michel Poggi is member of the Scientific Committee as well as the Organizing Committee of the colloquium MATHS A VENIR 2009 which promote the strong connections between mathematics on one side and industry, society and other sciences on the other side.

8.1.1. MONOLIX Group

Participant: Marc Lavielle.

The MONOLIX group chaired by Marc Lavielle and France Mentré (INSERM) is a multidisciplinary group, that exchanges and develops activities in the field of mixed effect models. It involves scientists with various backgrounds, interested both in the study and applications of these models academic statisticians (theoretical developments), researchers from INSERM (applications in pharmacology) and INRA (applications in agronomy, animal genetics and microbiology), and scientists from the medical faculty of Lyon-Sud University (applications in oncology). This multi-disciplinary group, born in October 2003, has been meeting every month.

Moreover, Marc Lavielle was responsible of an ANR project (projet blanc) on the MONOLIX software which started in 2006. This project was selected for an oral presentation at the Cité des Sciences in February 2009.

8.2. European actions

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

9. Dissemination

9.1. Scientific Communauty 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 "Journées de Probabilité 2009" in Poitiers.
- Marc Lavielle was invited speaker at JSM 2009 in Washington, Population PK 2009 meeting in London, PKUK 2009 in Birmingham, ACOP 2009 in Mystic (US), POPSIM Meeting in Copenhagen, SFdS 2009 meeting in Bordeaux, Biosafenet meeting in Ca Tron di Roncade (Italy), Workshop "Les OGM face aux nouveaux paradigmes de la biologie" in Paris, Atelier "Maths et Industrie" in Paris
- Marc Lavielle participated at the Table Ronde "Maths et Société", Colloque Maths à Venir, Paris
- Jean-Michel Poggi was invited speaker at TIES 2009 in Bologna, SIS 2009 in Pescara (Italy), SFC 2009 in Grenoble, Colloquium "Statistiques pour le traitement de l'image" in Université Paris 1.

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 coucil 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 scientific council of the CIMPA (Centre International de Mathématiques Pures et Appliquées).
- Marc Lavielle is member of the Guidance Committee of the ANR programm on Public Health.
- Pascal Massart is the head of the Department of Mathematics of University Paris-Sud.
- 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 is a member of the scientific committee of the forthcoming European Meeting of Staticians to be held in 2010 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

Year Publications

Doctoral Dissertations and Habilitation Theses

[1] J.-P. BAUDRY. *Sélection de modèle pour la classification non supervisée. Choix du nombre de classes.*, Université Paris-Sud, 2009, Ph. D. Thesis.

[2] V. VANDEWALLE. *Estimation et sélection de modèle en classification semi-supervisée.*, Université Lille 1, 2009, http://tel.archives-ouvertes.fr/tel-00447141/PDF/These.pdf, Ph. D. Thesis.

Articles in International Peer-Reviewed Journal

[3] A. ANTONIADIS, I. GIJBELS, J.-M. POGGI. *Smoothing non equispaced heavy noisy data with wavelets*, in "Statistica Sinica", vol. 19, n^{o} 4, 2009, p. 1371–1387.

[4] S. ARLOT. *Model selection by resampling penalization*, in "Electron. J. Statist.", vol. 3, 2009, p. 557–624 (electronic).

[5] S. ARLOT, G. BLANCHARD, É. ROQUAIN. *Some non-asymptotic results on resampling in high dimension, I: confidence regions*, in "Annals of Statistics", 2009, To appear.

[6] S. ARLOT, G. BLANCHARD, É. ROQUAIN. *Some non-asymptotic results on resampling in high dimension, II: multiple tests*, in "Annals of Statistics", 2009, To appear.

[7] S. ARLOT, P. MASSART. *Data-driven calibration of penalties for least-squares regression*, in "J. Mach. Learn. Res.", vol. 10, 2009, p. 245–279 (electronic), http://hal.archives-ouvertes.fr/hal-00243116/PDF/arlot08a.pdf.

[8] S. BOUCHERON, G. LUGOSI, P. MASSART. *On concentration of self-bounding functions*, in "Electronic Journal of Probability", vol. 14, 2009, p. 1884-1899.

[9] S. BOUCHERON, P. MASSART. *A high dimensional Wilks phenomenon*, in "Probability Theory and Related Fields", 2009, submitted.

[10] G. CELEUX, A. GRIMAUD, Y. LEFEBVRE, E. DE ROCQUIGNY. *Identifying variability in multivariate systems through linearised inverse methods*, in "Inverse Problems in Engineering", 2009, to appear.

[11] L. CUCALA, J.-M. MARIN, C. ROBERT, M. TITTERINGTION. *A Bayesian reassessment of nearest-neighbour classification*, in "J. Amer. Statist. Assoc.", vol. 104, n^{o} 485, 2009, p. 263–273.

[12] E. EGER, V. MICHEL, B. THIRION, A. AMADON, S. DEHAENE, A. KLEINSCHMIDT. *Deciphering Cortical Number Coding from Human Brain Activity Patterns*, in "Current Biology", vol. 19, 2009, p. 1608-1615.

[13] A. IACOBUCCI, J.-M. MARIN, C. ROBERT. *On variance stabilisation by double Rao-Blackwellisation*, in "Comput. Statist. Data Anal.", 2010, (to appear).

[14] F.-X. JOLLOIS, J.-M. POGGI, B. PORTIER. *Three non-linear statistical methods to analyze PM10 pollution in Rouen area*, in "CS-BIGS", vol. 3, n^{o} 1, 2009, p. 1–17.

[15] A. KNOPS, B. THIRION, E. HUBBARD, V. MICHEL, S. DEHAENE. *Mathematics as cortical recycling : Recruitment of an area involved in eye movements during mental arithmetic*, in "Science", vol. 324, 2009, p. 1583-1585.

[16] M. LEBRETON, S. JORGE, V. MICHEL, B. THIRION, M. PESSIGLIONE. *An automatic valuation system in the human brain : evidence from functional neuroimaging*, in "Neuron", 2009, (to appear).

[17] C. MAUGIS, G. CELEUX, M.-L. MARTIN-MAGNIETTE. *Variable selection for Clustering with Gaussian Mixture Models*, in "Biometrics", vol. 65, 2009, p. 701-709.

[18] C. MAUGIS, G. CELEUX, M.-L. MARTIN-MAGNIETTE. *Variable selection in model-based clustering: A general variable role modeling*, in "Computational Statistics and Data Analysis", vol. 53, 2009, p. 3872-3882.

[19] C. MAUGIS, B. MICHEL. *A non asymptotic penalized criterion for Gaussian mixture model selection*, in "ESAIM, P & S", 2009, To appear.

[20] J. PERRY, C. TER BRAAK, P. DIXON, J. DUAN, R. HAILS, A. HUESKEN, M. LAVIELLE, M. MARVIER, M. SCARDI, K. SCHMIDT, B. TOTHMERESZ, F. SCHAARSCHMIDT, H. VAN DER VOET. *Statistical aspects of environmental risk assessment of GM plants for effects on non-target organisms*, in "Environmental Biosafety Research", vol. 8, 2009, p. 65–78.

[21] R. SAVIC, M. LAVIELLE. *A new SAEM algorithm: Performance in Population Models for Count Data*, in "Journal of Pharmacokinetics and Pharmacodynamics", vol. 36, 2009, p. 367–379.

[22] N. VERZELEN. *Adaptive estimation of Stationary Gaussian fields*, in "Annals of Statistics", 2009, http://hal.inria.fr/inria-00353251/PDF/RR-6797.pdf, to appear.

[23] N. VERZELEN. *High-dimensional Gaussian model selection on a Gaussian design*, in "Ann. Inst. H. Poincaré Probab. Statist.", 2009, to appear.

[24] N. VERZELEN, F. VILLERS. *Goodness-of-fit Tests for high-dimensional Gaussian linear models*, in "Annals of Statistics", 2009, to appear.

[25] N. VERZELEN, F. VILLERS. *Tests for Gaussian graphical models*, in "Comput. Statist. Data Analysis", vol. 53, 2009, p. 1894–1905.

Articles in National Peer-Reviewed Journal

[26] C. MAUGIS, M.-L. MARTIN-MAGNIETTE, J.-P. TAMBY, J.-P. RENOU, A. LECHARNY, S. AUBOURG, G. CELEUX. *Sélection de variables pour la classification par mélanges gaussiens pour prédire la fonction des gènes orphelins*, in "MODULAD", vol. 40, 2009, p. 69-80.

Invited Conferences

[27] G. CELEUX. *Sélection de modèle pour la classification en présence d'une classification externe*, in "Journées de Probabilité 2009, Poitiers", June 2009.

[28] M. DELATTRE, M. LAVIELLE. *Estimation of Mixed Hidden Markov Models with SAEM. Application to daily seizures data*, in "PKUK Meeting, Birmingham, UK", November 2009.

[29] M. DELATTRE, M. LAVIELLE. *Estimation of Mixed Hidden Markov Models with SAEM. Application to daily seizures data*, in "ACOP Meeting, Mystic, US", October 2009.

[30] R. GENUER, J.-M. POGGI, C. TULEAU. *Random Forests: variable importance and variable selection*, in "SFC'2009, Grenoble", September 2009.

[31] R. GENUER, J.-M. POGGI, C. TULEAU. *Random Forests: variable importance and variable selection*, in "Colloquium Statistiques pour le traitement de l'image, Paris", January 2009.

[32] R. GENUER, J.-M. POGGI, C. TULEAU. *Variable Selection using Random Forests*, in "International Meeting on Statistical Methods for the Analysis of Large Data-Sets, Pescara", September 2009.

[33] F.-X. JOLLOIS, J.-M. POGGI, B. PORTIER. *Three non-linear statistical methods to analyze PM10 pollution in Rouen area*, in "TIES 2009 - the 20th Annual Conference of The International Environmetrics Society and GRASPA Conference 2009, Bologna", July 2009.

[34] M. LAVIELLE. *Analysing Population PK-PD Data with the SAEM Algorithm and MONOLIX*, in "Population PK-PD Meeting, London, UK", October 2009.

[35] M. LAVIELLE. *Problèmes statistiques de l'interprétation des essais toxicologiques*, in "Colloque "Les OGM face aux nouveaux paradigmes de la biologie", Paris", February 2009.

[36] M. LAVIELLE. *Some comments about the statistical methodology used for the analysis of toxicity tests*, in "Biosafenet Meeting, Ca Tron di Roncade, italy", January 2009.

[37] M. LAVIELLE, F. MENTRÉ. *Estimation et planification dans les modèles non linéaires à effets mixtes. Application à la dynamique du VIH sous traitement*, in "41èmes Journées de Statistique, Bordeaux", May 2009, http://hal.inria.fr/inria-00386791/PDF/p234.pdf.

[38] M. LAVIELLE, A. SAMSON, A. K. FIRMIN, F. MENTRÉ. *Parameter estimation of long-term HIV dynamic model in the COPHAR2 - ANRS 111 trial using MONOLIX*, in "Joint Statistical Meeting, Washington, US", August 2009.

[39] M. LAVIELLE, R. SAVIC. *Modeling odd-type data with MONOLIX*, in "POPSIM Meeting, Copenhagen, Danmark", September 2009.

[40] N. VERZELEN. *Data-driven neighborhood selection of a Gaussien field*, in "The 20th Annual Conference of The International Environmetrics Society, Bologna, Italy", July 2009.

International Peer-Reviewed Conference/Proceedings

[41] R. GENUER, V. MICHEL, E. EGER, B. THIRION. *Random Forests based feature selection for decoding fMRI data*, in "submitted to IEEE International Symposium on Biomedical Imaging", November 2009.

[42] M. KELLER, M. LAVIELLE, M. PERROT, A. ROCHE. *Anatomically Informed Bayesian Model Selection for fMRI Group Data Analysis*, in "12th MICCAI, London, U.K.", September 2009.

[43] C. MAUGIS. *Variable selection in model-based clustering: A general variable role modeling*, in "Classification and Data Analysis (CLADAG), Catania", September 2009.

[44] V. MICHEL, E. EGER, C. KERIBIN, B. THIRION. *Adaptive multi-class Bayesian sparse regression - An application to brain activity classification*, in "MICCAI'09 Workshop on Analysis of Functional Medical Images", 2009.

National Peer-Reviewed Conference/Proceedings

[45] V. VANDEWALLE. *A data-driven penalized criterion for Gaussian mixture model selection*, in "41èmes Journées de Statistique, Bordeaux", May 2009.

Workshops without Proceedings

[46] J.-P. BAUDRY. *Critères de sélection de modèles consistants pour la classification non supervisée.*, in "Séminaire MAP5, Paris", October 2009.

[47] J.-P. BAUDRY. *Sélection de modèle pour la classification non supervisée.*, in "Séminaire SAMOS-Paris I, Paris", October 2009.

[48] J.-P. BAUDRY, G. CELEUX. *Sélection de modèle pour la classification en présence d'une classification externe*, in "41èmes Journées de Statistique, Bordeaux", May 2009, http://hal.inria.fr/inria-00386620/PDF/p66.pdf.

[49] J.-P. BAUDRY, C. MAUGIS, B. MICHEL. *How to put the slope heuristics in practice*, in "Summer Working Group on Model-Based Clustering, Paris", July 2009, Poster.

[50] G. CELEUX. *Clustering Model Selection related to an External Classification*, in "Summer Working Group on Model-Based Clustering, Paris", July 2009.

[51] M. DELATTRE. *Application des modèles de Markov cachés à effets mixtes à des données d'épilepsie*, in "Journées du GDR Statistique et Santé, Paris", October 2009.

[52] M. DELATTRE, M. LAVIELLE. *Estimation of Mixed Hidden Markov Models with SAEM. Application to daily seizures data*, in "CLAPEM, Naiguata, Venezuela", November 2009.

[53] M. EL ANBARI, A. MKHADRI. *Regularization and variable selection via the Larcop*, in "41èmes Journées de Statistique, Bordeaux", May 2009.

[54] G. GOVAERT, G. CELEUX. *Block clustering and mixture models*, in "Summer Working Group on Model-Based Clustering, Paris", July 2009.

[55] C. MAUGIS, G. CELEUX, M.-L. MARTIN-MAGNIETTE. *Sélection de variables pour la classification non supervisée par mélanges gaussiens et pour l'analyse discriminante gaussienne*, in "41èmes Journées de Statistique, Bordeaux", May 2009.

[56] C. MAUGIS. *Variable selection for model-based clustering and discriminant analysis*, in "Summer Working Group on Model-Based Clustering, Paris", July 2009.

[57] A. PASANISI, C. ROERO, G. CELEUX, E. RÉMY. *Quelques considérations sur l'utilisation pratique des modèles discrets de survie en fiabilité industrielle*, in "41èmes Journées de Statistique, Bordeaux", May 2009, http://hal.inria.fr/inria-00386574/PDF/p20.pdf.

[58] V. VANDEWALLE. *Model selection in semi-supervised classification*, in "Working Group on Model-Based Clustering Summer Session, Paris", July 2009.

[59] V. VANDEWALLE. *Sélection de modèles en classification semi-supervisée*, in "Groupe de travail AgroParisTech-Paris Descartes-select, Paris", March 2009.

Scientific Books (or Scientific Book chapters)

[60] G. CELEUX. *Discriminant Analysis*, in "Data Analysis", John Wiley, 2009, p. 181-214.

Research Reports

[61] Y. AUFFRAY, P. BARBILLON. *Conditionally positive definite kernels : theoretical contribution, application to interpolation and approximation*, n^{o} RR-6835, Institut National de Recherche en Informatique et Automatique, 2009, http://hal.inria.fr/inria-00359944/PDF/RR-6835.pdf, Technical report.

[62] Y. AUFFRAY, P. BARBILLON, J.-M. MARIN. *Maximin Design on non-hypercube domain and Kernel Interpolation*, Institut National de Recherche en Informatique et Automatique, 2009, Technical report.

[63] P. BARBILLON, G. CELEUX, A. GRIMAUD, Y. LEFEBVRE, E. DE ROCQUIGNY. *Non linear methods for inverse statistical problems*, Institut National de Recherche en Informatique et Automatique, 2009, http://hal.inria.fr/inria-00441967/PDF/RR-7156.pdf, Technical report.

[64] J.-P. BAUDRY, A. E. RAFTERY, G. CELEUX, K. LO, R. GOTTARDO. *Combining Mixture Components for Clustering*, n^{o} 6644, Institut National de Recherche en Informatique et Automatique, 2009, Research Report.

[65] C. GIRAUD, S. HUET, N. VERZELEN. *Graph selection with GGMselect*, hal-00401550, 2009, http://hal.archives-ouvertes.fr/hal-00401550/en/, Technical report.

[66] C. KERIBIN. *Les méthodes bayésiennes variationnelles et leur application en neuroimagerie: une étude de l'existant*, n^{o} RR-7091, Institut National de Recherche en Informatique et Automatique, 2009, http://hal.inria.fr/inria-00430289/PDF/RR-7091.pdf, Technical report.

[67] C. MAUGIS, B. MICHEL. *Slope heuristics for variable selection and clustering via Gaussian mixtures*, n^{o} RR-6550, Institut National de Recherche en Informatique et Automatique, 2009, http://hal.inria.fr/docs/00/28/50/32/PDF/RR-6550.pdf, Technical report.

[68] M. MISITI, Y. MISITI, G. OPPENHEIM, J.-M. POGGI. *Stratégie divisive pour la prévision par désagrégation - Parallélisation et effet de la taille des données*, Rapport EDF (60 pages), 2009, Technical report.

[69] N. VERZELEN. *Data-driven neighborhood selection of a Gaussian field*, n^{o} 6798, Institut National de Recherche en Informatique et Automatique, 2009, http://hal.inria.fr/inria-00353260/PDF/RR-6798.pdf, Technical report.

Scientific Popularization

[70] V. VANDEWALLE. *Les modèles de mélange, un outil utile pour la classification semi-supervisée*, in "La revue MODULAD", vol. 40, 2009, p. 121-145.