Contribution to Statistics and Data Science for industrial applications. From neural networks to sparse linear models.

Mathilde Mougeot

Mémoire présenté à l'Université Paris-Diderot

par

Mathilde MOUGEOT

Pour l'obtention d'une

Habilitation à Diriger des Recherches

Contribution to Statistics and Data Science

for industrial applications.

From neural networks to sparse linear models.

Soutenue le 7 Décembre 2015 devant le jury composé de MM.

Alfred	**Hero**	U. Michigan	*Rapporteur*
Stéphane	**Boucheron**	U. Paris Diderot	*Rapporteur*
Pascal	**Massart**	U. Paris Sud	*Rapporteur*
Sylvain	**Sardy**	U. Genève	*Rapporteur*
Stéphane	**Canu**	INSA Rouen	*Examinateur*
Dominique	**Picard**	U. Paris Diderot	*Examinateur*
Gilles	**Stoltz**	HEC	*Examinateur*
Nicolas	**Vayatis**	ENS Cachan	*Examinateur*
Michèle	**Sebag**	U. Paris Sud	*Invitée*

Laboratoire de Probabilités et Modèles Aléatoires - UMR 7599

Je tiens à remercier :

- Stéphane Boucheron, Alfred Hero, Pascal Massart et Sylvain Sardy,

  Rapporteurs de ce mémoire, pour leurs points de vue critiques et leurs retours précieux sur ces travaux, ainsi que

- Stéphane Canu, Dominique Picard, Michèle Sebag, Gilles Stoltz, Nicolas Vayatis,

  qui ont sans hésiter accepté de faire partie de mon Jury.

Ce travail est le fruit de collaborations successives et n'aurait pas pu voir le jour sans la contribution de mes co-auteurs, que je remercie: Aurélie, Antoine, Cristina, Charles, Daniel, Dominique, Gérard, Gilles & Gilles, Jérôme, Julien, Karine, Laurence, Lucette, Mina, Nicolas, Olivier, Robert, Stephan, Sylvain et Vincent.

Aujourd'hui en écrivant ces lignes, je tiens également à remercier les collègues que je ne peux pas tous citer et qui ont indirectement enrichi les réflexions et le travail scientifique à l'origine de ce mémoire, collègues croisés au laboratoire d'Orsay, à Thales, au DMA de l'ENS, au CMLA de l'ENS Cachan, à Modal'X de l'Université Paris-Ouest-Nanterre-La Défense, à Miriad Technologies, et au LPMA de l'université Paris Diderot et de l'UPMC.

Merci.

Contents

I Scientific Overview **v**

Career **vii**

Publications **xi**
Research . . . xi
Teaching . . . xiii
Softwares . . . xiv

Projects **xv**
Academic Projects . . . xv
European Projects . . . xv
Academic/Industrial collaborations . . . xv
PHD thesis supervision . . . xvi

II Scientific work **1**

Introduction **3**

1 High dimensional linear models **5**
1.1 ℓ_1 penalization . . . 6
1.2 Thresholdings . . . 9
1.2.1 Learning out of Leaders . . . 9
1.2.2 Main theoretical results . . . 12
1.2.3 Heuristics for calibration . . . 15
1.3 Numerical experiments . . . 16
1.3.1 Ultrahigh dimension . . . 17
1.3.2 Two-step procedures . . . 17
1.3.3 Beyond theoretical assumptions . . . 17
1.4 Conclusions . . . 18

2 From functional regression to electrical consumption forecast **19**
2.1 Functional regression and HD linear models . . . 21
2.2 Mining the curves using sparse approximation . . . 22

2.2.1 Choice of a generic dictionary . . . 22
2.2.2 Mining and clustering to define adaptive patterns of consumption . . . 24
2.2.3 Patterns of consumption as endogenous variable . . . 24
2.3 Modeling . . . 24
2.3.1 Sparse model and adaptive dictionary for modeling the curves . . . 25
2.4 Forecasting . . . 25
2.4.1 The experts . . . 26
2.4.2 Aggregation . . . 26
2.4.3 Performances . . . 27
2.5 Software . . . 28
2.6 Conclusions . . . 28

3 Grouping variables for high dimensional linear models 29
3.1 Group penalization . . . 30
3.2 Thresholdings with groups . . . 30
3.2.1 Learning Out of Leaders algorithm with groups . . . 31
3.2.2 Main theoretical results . . . 33
3.2.3 Grouping versus no grouping . . . 34
3.3 Data driven strategies to built relevant groups . . . 35
3.4 Numerical experiments and applications . . . 37
3.4.1 Experimental design . . . 37
3.4.2 Catching feature with the grouping strategy . . . 38
3.5 Software . . . 41
3.6 Conclusions . . . 41

4 Industrialization of statistical or machine learning algorithms 43
4.1 From research to development: step by step . . . 43
4.1.1 Before the project . . . 44
4.1.2 Proof of concept (POC) . . . 44
4.1.3 Pilot . . . 45
4.1.4 Industrial software . . . 46
4.2 Embedding R&D in Software . . . 46
4.2.1 Monitoring Energy Performance of compressors . . . 46
4.2.2 Automated Diagnosis for Helicopter Engines and Rotating parts . . . 49
4.3 Health equipment monitoring with predictive modeling . . . 51
4.3.1 Supervised classification vs anomaly detection . . . 51
4.3.2 Expert knowledge vs knowledge extraction . . . 51
4.3.3 Robustness of an automatic decision in an operational environment . . . 52
4.4 Conclusions . . . 52

Perspectives 53

Bibliography 54

III Annexes **61**

Industrial Technical reports **63**

Part I

Scientific Overview

Scientific career overview

I began my academic career in 1992 at Paris X University[1] as an associate professor. At that time, I was working on artificial neural networks for the modeling of biological networks [A2], [C1-C3] and for image compression [A1-A3]. I was rapidly involved in collaborative projects between the Centre de Mathématiques et de Leurs Applications (CMLA) at Ecole Normale Supérieure de Cachan (ENSC) and industrial partners, and became very interested in statistical and machine learning for the monitoring of industrial processes. Neural networks associated with more classical signal processing techniques were introduced early in 1994 for the health monitoring of the Vulcain engine of the Ariane rocket in an initial collaboration with the CNES and SEP [Anx: Miriad Tec. Reports].

In October 1999, I left my position at Paris X University to participate in the creation of Miriad Technologies, a private company co-founded by Robert Azencott. The creation of the company was an opportunity to test my scientific experience, developed through previous industrial partnerships. At that time, it was particularly innovative and challenging to propose the development of ad hoc solutions with original mathematical concepts implemented in light software. During this period, I held a Research and Development position and was in charge of the "Proof of concepts" (POC) division: given an industrial problem, my work was, first, to develop a methodology using statistics or machine learning tools, and to then implement the solution to evaluate the performances on raw operational data. The originality of our approach was to combine, very soon, machine-learning techniques with more classical methods such as statistics or signal processing to answer to industrial needs. We developed our own software, called Miriad Process, for Rapid Application Development of POCs. In order to answer frequently asked problems, we created new tools to mine industrial data such as an innovative method based on a mutual information ratio and entropy to be able to rank the factors responsible for a quality defect. This method made it possible to detect non-linear relationships and was also used for variable selection before regression models [C6, E2]. An accomplishment that I am particularly proud of was the development of a method able to automatically diagnose over-consumption for industrial compressors based on operational data analysis. During a six-month field validation, this method showed excellent performances for detection. For the following two years, I worked as a project manager to supervise the development of industrial software, to monitor the first industrial applications, and to implement the software in the United States for Air Liquid America [C5]. Until now, this software had been used to monitor on-line working compressors at the Operational Control Command in Houston, Texas (USA).

[1] today Paris Ouest Nanterre La Défense University

Most of the projects at Miriad were developed under non-disclosure agreements and were not published in academic journals. Technical reports were, however, systematically written for each POC and delivered to the concerned clients or partners. For six years, I had the incredible opportunity of working on various applications for more than twenty different industrial partners [Anx: Miriad Tec. Reports] and of participating in the challenging adventure of a start-up that aimed at selling innovative software solutions based on statistical or machine-learning methodologies.

Because I wanted to spend more time on research than on development, I chose to return to my academic position at Paris X in October 2005, when Miriad Technologies turned to focus exclusively on business applications. During the following three years, I took part in the scientific and administrative work of the European projects ADHER (Automated Diagnosis for Helicopter Engines and Rotating parts, Eu 030907) and Innotex (INNOvation within the TEXtile manufacturing lines in Europe, Eu 030312) that were transferred from Miriad Technologies to the CMLA at ENSC, my former lab, where I naturally returned to work. With J. Wang, we made significant efforts to develop a method based on predictive modeling to monitor helicopter engines using a massive data analysis of vibration and contextual flight data. We delivered a corresponding software used by RSL and Eurocopter and the method has received excellent feedback from the field [E7, E11]. Working for data mining applications, I was involved in the TRACE European project (TRAffic Causation Analysis in Europe), mining one of the largest databases for car accidents in Europe, the German In-Depth Accident Study data base (GIDAS). Within the framework of this project, we introduced a greedy algorithm based on a mutual information ratio to quantify the root causes of car accidents [E2, C8].

In October 2009, I was offered an associate professor position at Paris Diderot University in the Laboratoire de Probabilités et Modèles Aléatoires (LPMA). This environment provided me the opportunity to discover and to investigate new research directions based on mathematical statistics. With K. Tribouley, we introduced a test procedure to compare tail indices and applied this procedure to compare the risk behaviors of a panel of different financial data [A7]. Challenging ℓ_1-penalization methods, a very effective procedure with no optimization, were developed for prediction in high dimension, in collaboration with D. Picard [A8, A12], and extended to the case of grouping variables [A9]. This model was used in the functional regression framework to model and forecast intra-day electrical consumption signals, within the framework of a scientific partnership with RTE[2] [A11, A13]. A software was also developed and delivered to RTE. During the same time, collaborations with the CEA were initiated, with S. Delattre, to develop dedicated estimation end interpolation tools for thermodynamic experiments [C9].

My former experience in the R&D of industrial applications strongly influences my research today. My aim is to work on new statistical or machine learning methods but with an operational purpose always in mind. My extensive experience both in research and in development helped me to create a link between academic skills and industrial issues. The need for Data Science and predictive modeling in private companies has increased in recent years and many opportunities in this area have emerged for public laboratories. However, I find that it is always a challenging task to successfully develop innovative scientific research and to respond, at the same time, to a real operational need. From my point of view, specific structures are still needed to make

[2]Résau de Transport d' Electricité

the bridge between the two worlds and to transform statistical innovative methods into useful operational software. Since I have returned to my associate professor/researcher position at the university, I have made significant efforts to develop original scientific collaborations within the academic framework with Snecma (2009), Air Liquide (2010), RTE (2011-2015) and the CEA (2012-2015).

Since 2014, I supervise the PHD thesis of Mina Abdel-Sayed, in collaboration with G. Faÿ, Centrale Supelec, and SAFRAN for the detection of potential failures in high dimensional spectrograms [C10].

Teaching activities

At Paris X, I introduced the use of software computation and statistical software very early in 1992 to teach statistics in the Mathematics Department. The lectures, created for and addressed to Master students, were quite innovative at that time and I was invited to present that experience in international workshops [T1-T3]. I deeply rely today on the practical experience I acquired through my previous industrial collaborations for my lectures. The pedagogical line I have built over the last years prevents me from introducing statistical theory without presenting any application perspectives or practical software applications. Starting with any type of numerical operational data, I aim at interacting with the students about the practical functionalities that can be implemented using statistics or machine-learning methodologies. I provide, at the same time, the related benefits or drawbacks of uses in potential applications [T4]. Due to the recent emergence of various open databases, this approach is much easier to carry out today. I use a similar approach for undergraduate and Master students, obviously with differences depending on their academic background. Consistent with this pedagogical point of view, I was asked to give lectures at the Master level at Paris Diderot University (2009-.), at the Ecole Centrale de Paris with N. Vayatis (2008-.) and at the Ecole Normale Supérieure with C. Zalc (2010-.). At ENS, I developed a dedicated course for using statistics in the research of students working in other divisions (geographic, social and human sciences).

Tremendous and rapid progress in the implementation of automatic decision making processes can be easily made within the industry by introducing dedicated and appropriate teaching linked to the presentation of real operational applications and their associated added value. To address this specific need, the Snecma Company asked me to create a three-day data mining and predictive modeling course to help aeronautic engineers to apply statistics and data mining tools to their research activities, which took place in 2013, and which should soon be extended to other divisions.

Publications

Peer-reviewed journal papers

A13 M. Mougeot, D. Picard, V. Lefieux, L. Maillard-Teyssier. (2015) *Modeling and Stochastic Learning for Forecasting in High Dimension.* Springer Lecture Notes in Statistics, , p161-182.

A12 M. Mougeot, D. Picard, K. Tribouley. (2014) *LOL selection in high dimension.* Computational Statistics & Data Analysis p 743-757.

A11 M. Mougeot, D. Picard, K. Tribouley. (2013) *Sparse approximation and fit of intraday load curves in a high dimensional framework.* Advances in Adaptive Data Analysis p 1-23.

A10 M. Mougeot, D. Picard, K. Tribouley. (2013) *Grouping Strategies and Thresholding for High Dimensional Linear Models rejoinder.* Journal of Statistical Planning and Inference 143, p 1457-1465.

A9 M. Mougeot, D. Picard, K. Tribouley. (2013) *Grouping Strategies and Thresholding for High Dimensional Linear Models, with discussion* Journal of Statistical Planning and Inference 143, p 1417-1438.

A8 M. Mougeot, D. Picard, K. Tribouley (2012) *Learning Out of Leaders.* J. R. Stat. Soc. Ser. B Stat. Methodol. p 1–39.

A7 M. Mougeot, K. Tribouley (2010) *Procedure of test to compare the tail indices* Annals of the Institute of Statistical Mathematics p 383–412.

A6 J.J. Stirnemann, M. Mougeot, F. Proulx, B. Nasr B. M. Essaoui, J.C. Fouron, Y. Ville. (2010) *Profiling fetal cardiac function in twin to twin transfusion syndrome.* Ultrasound Obstet Gynecol, 35 p 19–27.

A5 G. Kerkyachariana, M. Mougeot, D. Picard, K. Tribouley. (2009) *Learning Out of Leaders.* Multiscale, Nonlinear and Adaptive Approximation (2009), p 293–322.

A4 C. Butucea, M. Mougeot, K. Tribouley. (2007) *Functional approach for excess mass estimation in the density model.* Electronic Journal of Statistics, p 449–472.

A3 M. Mougeot, R. Azencott, B. Angeniol. (1991) *Image compression with back propagation: improvement of the visual restoration using different cost functions.* Neural Networks, 4, 4, p 467–476.

A2 M. Mougeot, (1991) *Self-organization of orientation selection cells and intracortical connections in the visual cortex.* Pattern recognition and neural networks, vol 23, p 63-73.

A1 M. Mougeot, R. Azencott, B. Angéniol. (1989) *A study of image compression with backpropagation.* Neuro Computing, Algorithms, Architectures and Applications, ASI series. Series F: computer and system sciences.

Main Peer-reviewed conference papers

C10 M. Abdel-Sayed, D. Duclos, G. Faÿ, J. Lacaille, M. Mougeot (2015) NMF-based decomposition for anomaly detection applied to vibration analysis. Proceedings of the 10th International Conference on Condition Monitoring and Machinery Failure Prevention Technologies, June 2015.

C9 E. Fraizier, S. Delattre, M. Mougeot, Ph. Faure, G. Roy, G. F. Poggi (2012) *A new probabilistic tool for the determination and optimization of multiphase Equation Of State parameters: application to tin.* DYMAT - 10th International Conference on the Mechanical and Physical behavious of Materials under Dynamic Loading Freiburg, Germany, September 2nd-7th, 2012. ISBN:978-2-7598-0757-4.

C8 M. Mougeot, R. Azencott, R. (2011) *Traffic safety: non-linear causation for injury severity.* Safety and Security Engineering IV (2011), WIT Press, p 241-252.

C7 M. Mougeot, K. Tribouley. (2008) *Goodness-of-fit independance test based on Excess Mass properties for copula.* Proceedings Multimodality and related topics.

C6 M. Mougeot, R. Azencott (2008) *Information theoretical methods dedicated to accidents analysis for GIDAS database.* European Symposium on Accident Research, Hannovre.

C5 O. Cadet, C. Harper, M. Mougeot (2005) *Monitoring Energy Performance of Compressors with an innovative auto-adaptive approach.* Instrumentation System and Automation -ISA- Chicago.

C4 M. Mougeot. (1997) *Synchronization and Oscillations in the visual cortex: a stochastic model using a spike memory term.* European Symposium on Artificial Neural Networks, Bruges April.

C3 M. Mougeot (1996) *Biological Model of synchronization in the visual cortex.* World Congress on Neural Networks, San Diego.

C2 M. Mougeot, R. Azencott (1991) *Unsupervised learning for the visual cortex (layer IV): model and simulations.* Internal Joint Conference on Neural Network, IJCNN (1991).

C1 M. Mougeot, R. Barrow (1990) *From Static to dynamic image compression.* International Conference on Neural Networks-ICNN-, Paris.

Main European reports

E11 ADHER (2009) *Automated Diagnosis for Helicopter Engines and Rotating parts* Publishable final activity report (2009), ADHER FP6/Aeronautics project AST5-CT-2006-030907.

E10 M. Mougeot (2008) *Self-learning modules for off-line optimization of quality by adjustments of control parameters.* Deliverable D2.3., INNOTEX european project 030312.

E9 M. Mougeot (2008) *Detection of quality risks based on process sensors recordings.* Deliverable D2.2. INNOTEX European project 030312

E8 R. Azencott, M. Mougeot, J. Wang (2008) *Auto adaptive elimination of degraded sensor inputs. Specifications and test results.* Deliverable Work package WP 3.4.ADHER FP6/Aeronautics project AST5-CT-2006-030907. Restricted.

E7 R. Azencott, M. Mougeot, J. Wang. (2008) *Automated Diagnosis for Helicopter Engines and Rotating parts.* Deliverable Work package WP 3.3. ADHER FP6/Aeronautics project AST5-CT-2006-030907. Restricted.

E6 M. Mougeot (2007) *Root causes Diagnosis component.* Deliverable D2.1. (2007) INNOTEX European project 030312.

E5 R. Azencott, M. Mougeot, J. Wang (2007) *Impact Analysis of contextual variables on vibrations.* Deliverable Work package WP 3.2. ADHER FP6/Aeronautics project AST5-CT-2006-030907. Restricted.

E4 R. Azencott, M. Mougeot, J. Wang (2007) *State of the art for off the shelf vibrations diagnosis softwares.* Deliverable Work package WP 3.1. ADHER FP6/Aeronautics project AST5-CT-2006-030907. Restricted.

E3 M. Mougeot, R. Azencott (2007) *Application to informations theoretic methods to GIDAS data base.* Deliverable 3.3-2. (2007) TRACE European project 027763.

E2 R. Azencott, J.P. Kreiss, M. Mougeot, P. Pastor, M. Pfeiffer, S. Siebert, T. Zangmeister. (2007) *Analysis Methods for Accident Causation Studies.* TRACE European project No. 027763.

E1 M. Mougeot, R. Azencott (2006) *Information theoretic methods and algorithms for accident causation analysis.* Deliverable 3.3-2. TRACE European project 027763.

Patent

P1 M. Abdel-Sayed, D. Duclos, J. Lacaille, G. Faÿ, M. Mougeot (14 avril 2015) Procédé de détection d'anomalie de turbomachine par analyse vibratoire automatisée. INPI, submission No. 1000288783.

Miriad Technical reports

see Annex/Miriad Technical reports

Thesis

M2 M. Mougeot. (1992) *Modèles connexionnistes appliqués à la compression d'images et à l'auto organisation du système visuel des mammifères.* Thèse de doctorat, Université ParisXI-Orsay.

M1 M. Mougeot. (1987) *Logiciel de calcul de vents réels pour la course de l'Admiral' Cup à l'aide de techniques neuronales.* Mémoire de Master, Université Paris VI.

Teaching papers

T4 M. Mougeot, G. Stoltz, (2016) *La statistique Connectée.* Journal Statistique et Enseignement, SFDS.

T4 M. Mougeot. (2007) *Impact of new technologies for Teaching statistics.* Proceedings of the International Statistical Institute, ISI 56^{th}, 22 -29 août, Lisbonne.

T3 L. Carter, M. Mougeot. (1998) *Use of Excel in a first course in Statistics for Mathematical Students.* ICATS-5, The first International Conference on Teaching Statistics, Singapore.

T2 L. Carter, M. Mougeot (1994) *Simulations to illustrate results in theoretical statistics.* Proceedings of four International Conference on teachings Statistics Marrakech, july.

T1 L. Carter, M. Mougeot. (1993) *Enseignement des statistiques assisté par ordinateur pour économistes.* Actes des XXV journées de statistiques de l'ASU, Vannes 24-28 Mai.

Softwares

S6 Group-LOL: Gouping Learning Out Of Leaders, 2013.
Diffusion: public with [A9].

S5 RTE: Sparse Modeling and Forecast for intra day load curves, 2012.
Diffusion: RTE.

S4 ADHER: Automated Diagnosis for Helicopter Engines and Rotating parts, 2009.
Diffusion: partners of the ADHER project.

S3 INNOTEX: INNOvation within the TEXtile manufacturing lines in Europe, 2010.
Diffusion: partners of the INNOTEX project.

S2 TRACE: TRAffic Causation Analysis in Europe, 2008.
Diffusion: partners of the TRACE project.

S1 SPEC+: Surveillance de la performance énergétique des compresseurs, 2005.
Diffusion: Air Liquide America.

Projects

Academic Projects

ANR (2014-2017) FOREWER: Modeling, FOrecasting and Risk Evaluation of Wind Energy pRoduction".
Project leader: P. Tankov, LPMA, Paris Diderot University.
Leader of the task "From resource distribution to power production": modeling the signals.

PGM0 (2014-2015). ALLO: Active Learning, Links with Optimization.
Programme Gaspard Monge pour l'Optimisation et la recherche opérationnelle.
Project leader: N. Vayatis, ENS Cachan.

European Projects

TRACE (2006-2008) TRAffic Causation Analysis in Europe, Eu. 027763
8 Parners. I was involved in the ENSC[3] tasks, 6 men months over 3 years.
http://www.trace-project.org/

ADHER (2008-2009) Automated Diagnosis for Helicopter Engines and Rotating parts, Eu 030907
5 Partners: Eurocopter, RSL Electronics,Cardiff University, Patras University, ENSC. I was involved in the ENSC tasks, 30 men months for 2 years, coordination of SP3.

INNOTEX (2008-2010) INNOvation within the TEXtile manufacturing lines in Europe. Eu. 030312,
26 Partners. I was involved in the ENSC tasks, 25 men months over 3 years.

Academic/Industrial collaborations

Collaborations initiated between Universities and industrial partners:

AIR LIQUIDE/ Université Paris Diderot (2010-2011).
"Aide statistique à la surveillance de procédés".

CEA/ Université Paris Diderot (2014-2015)
"Interpolation de l'énergie avec contraintes thermodynamiques", with S. Delattre.

CEA/ Université Paris Diderot (2013-2014)
"Estimation et Interpolation de l'entropie avec contraintes de monotonie", with S. Delattre.

[3]ENSC: Ecole Normale Supérieure de Cachan

CEA/ Université Paris Diderot (2011-2012)
"Calibration de constantes physiques: modèles et procédure d'estimation" with S. Delattre.

RTE/ Université Paris Diderot (2015-2016)
"Prévision fonctionnelle sur 48H des signaux de consommation électrique" with D. Picard.

RTE/ Université Paris Diderot (2014-2015)
Segmentation non supervisée de signaux météorologiques with D. Picard.

RTE/ Université Paris Diderot (2012-2013)
"Prévision des signaux de consommation électrique" with D. Picard, K. Tribouley.

RTE/ Université Paris Diderot (2010-2011)
"Modélisation des signaux de consommation électrique", with D. Picard.

SNECMA/ Université Paris Ouest Nanterre La Défense (2009-2010)
"Algorithme de lissage de champs de compresseurs" with T. Jeantheau, K. Tribouley.

PHD thesis supervision

Co-supervision of Mina Abdel-Sayed with G. Faÿ, Ecole Centrale Supélec , "Représentations optimisées et analyse automatique de signatures vibratoires" (2014-.). Thèse CIFRE, en collaboration avec les sociétés SAFRAN et SNECMA.

Part II

Scientific work

Introduction

As mentioned by Breiman [2001] fifteen years ago, there are "Two cultures in the use of statistical modeling to reach conclusions from the data". The first culture found its roots in the mathematical and statistical community and is mostly interested by a theoretical research in statistics "exploring inference, hypotheses testing and asymptotic" and today "oracle inequalities". The second culture emerged nearly fifty years ago initially driven by the computer science and the electrical engineering community. At the opposite of the first one, this culture deals with algorithmic modeling to catch relationships between inputs and outputs. Model validation is exclusively performed by predictive accuracy, measured on raw data [Vapnik, 1982]. For a long time, these two cultures have worked beside each other. Besse et al. [2001] have provided an introduction to *D*ata Mining in the form of a reflection about the interactions between Data processing and Statistics collaborating in the analysis of large sets of data. From both sides, practical and theoretical interrogations have progressively emerged such as:

-How is it possible to use in practice the conclusions of a theorem which "assume that the data are generated by the following model...? [Breiman, 2001]

-Are the theoretical assumptions, which explicitly guarantee asymptotic convergence or oracle inequalities, computable? in a given time?

-Could we propose heuristics to calibrate theoretical parameters, to link theory and practice?

-Should we trust the prediction results provided by a "black box" calibrated on data without theoretical guarantee on the underlaying statistical distribution ?

-Can we industrialize softwares based on statistical or machine learning algorithms?

Since the beginning of my career, I have felt the need to mix the two cultures and my successive works have progressively formed my views about algorithmic and theoretical modeling. This manuscript presents selected examples of my contribution to possible answers to the previous questions.

Inspired by the practical use of High Dimensional (HD) linear models, the first chapter of this manuscript presents an alternative procedure to the ℓ_1 penalization method (LASSO) called Learning Out of Leaders (LOL), developed in collaboration with G. Kerkyachariana, D. Picard and K. Tribouley. This procedure is simply based on thresholding to estimate the coefficients of a sparse model, and does not need any optimization step. As Restricted Isometric Properties (RIP) or restricted eigenvalues conditions [Bühlmann and Van De Geer, 2011] are needed to prove oracle inequalities for the LASSO, the theoretical behavior of the LOL procedure is driven by a simple index, called the empirical coherence, which can be easily computed from the data. The consistency relies on exponential bounds, leading to minimax and adaptive results for a wide class of sparse parameters, with (quasi) no restriction on the number of regressors. Benefits of using the LOL procedure are particularly apparent in ultra large dimensions when the computational optimization cost of the ℓ_1 procedures may be a potential hurdle. To make a bridge between the theoretical setting and the practical use of the procedure, we have introduced a heuristic, to be able to compute data driven thresholds. Implementation for large experimental designs or for real applications demonstrates the ability of the procedure to yield adapted calibration values.

In 2011, V. Lefieux and L. Teyssier-Maillard from the Research and Development department of RTE[4] asked us a practical question: *"Is it possible to built forecast models in the electricity consumption field which would rely on very few parameters and would be easy to calibrate -without the need of human expertise- and which, at the same time, would show good performances?"*. Theoretical and practical advances were required in order to meet RTE' s demands and chapter II presents an application of the functional use of the HD regression model to approximate and to forecast the intra day load curves using sparse linear models. Data mining on an historical set of data was first performed to catch relevant functional features, which appeared to be essential to built sparse models. The final forecast was computed using an aggregation of different forecast experts. A software is currently running at RTE which investigates the performances of the methodology on new operational data.

In the hope of taking advantages of a prior knowledge, group structures may be introduced in a regression model. Chapter III shows that the LOL procedure can be easily extended to estimate or discard groups of coefficients with a theoretical behavior similar to the Group Lasso, but again with a cheaper computational cost as in Chapter I. When no prior relation is imposed on the design, a major and difficult question is how to infer such a good structure automatically from the data, in order to improve the prediction performances. To optimize the rate of convergence of the Group LOL procedure, with D. Picard and K. Tribouley, we proposed a Boosting Grouping strategy to reorganize the initial predictors into relevant groups. This data driven strategy addressed the question of choosing the number of groups, as the distribution of the initial predictors across the groups by scattering then gathering the variables. A software package of the Group LOL algorithm has been developed and used on an experimental neuroscience data set by [Mairal and Yu, 2013]. Complementary numerical experiments showed the practical benefits of the Boosting Grouping strategy.

At the opposite of mathematical statistics, a Data Science project does not start with mathematical assumptions but with raw data and operational requirements as recently underlined by Wickham [2014]. The final statistical model is often the "tip of the iceberg" and can not even emerge if a large part of the work is not previously devoted to exploration, cleaning, pre-processing... Based on my 6 year experiences at Miriad Technologies start-up and my long experience in academic/industrial collaborations since 1999, the last chapter focuses on embedding statistical or machine learning algorithms within industrial software applications. For illustration, two success stories of monitoring, based on predictive modeling, are presented, which started as proof of concepts and finally led to softwares, deployed in the industry for daily uses. In the Miriad Technologies framework, we developed, the SPEC software, which provides a methodology to monitor and to diagnose overconsumption for large compressors working in the industry. Within the European project ADHER, we developed a software to provide Automatic Diagnosis for Helicopter Engines and Rotating parts.

[4]This work has been realized thanks to contractual collaborations with Réseau transport électrique (RTE) from 2010 to 2015

Chapter 1

High dimensional linear models

*T*he work presented in this chapter has been performed through a collaboration with G. Kerkyachariana, D. Picard and K. Tribouley. It has been published in:

A12 M. Mougeot, D. Picard, K. Tribouley. (2014) *LOL selection in high dimension.* Computational Statistics & Data Analysis p 743-757.

A8 M. Mougeot, D. Picard, K. Tribouley (2012) *Learning Out of Leaders.* J. R. Stat. Soc. Ser. B Stat. Methodol. p 1–39.

A5 G. Kerkyachariana, M. Mougeot, D. Picard, K. Tribouley. (2009) *Learning Out of Leaders.* Multiscale, Nonlinear and Adaptive Approximation (2009), p 293–322.

—

Today, it is usual to observe data sets with more variables than the number of observations. For example, in aeronautics, high dimensional spectrogram images are generated for few engine tests [C10] and in genomic, gene expression are often studied given a huge number of initial genes compared to a relatively low number of observations [Bickel et al., 2009a]. High Dimensional (HD) models have today a lot of practical applications.
The linear model:
A simple yet very useful model is the linear model:

$$Y_i = X_i\beta + \varepsilon_i \tag{1.1}$$

where Y_i is the target variable and $X_i = (X_{i,1}, \dots, X_{i,p})$ are the p covariates (where the constant parameter is included, $X_{i,1} = 1$) and ε_i is a non observed random Gaussian error, $N(0, \sigma^2)$ with $\{\varepsilon_i, 1 \leq i \leq n\}$ independent. The unknown parameters $\beta \in \mathbb{R}^p$ are frequently estimated by minimizing the ℓ_2 norm (which may be normalized by the number n of observations):

$$\hat{\beta}^{OLS} = \min_{\beta} \frac{1}{n}\|Y - X\beta\|^2_{\ell_2}$$

where Y is a $(n, 1)$ vector containing the target observations, X is the (n, p) design matrix and β is the $(p, 1)$ coefficient vector.

Various objectives can lead to analyze data using a linear model:

- regarding data mining applications, **information** about the "most" significant coefficients, brings knowledge about the important features which linearly explain the target [Guyon and Elisseeff, 2003]. For example, for the European TRACE project, (generalized) linear models were, in particular, studied to understand the TRaffic Accident causations in Germany [E2].
- Regarding **predictive modeling** applications, once the coefficients $\hat{\beta}$ are estimated, and given new values of the co variables x_{new}, the target prediction, $\hat{y}_{new} = y_{new}\hat{\beta}$, may be computed and used, for example, for forecasting purpose as to predict the intra day load curves in the electricity area [A13].
- knowing the observed value of the target, y_{new} may help for the monitoring or diagnosis of this new observation based on the residual analysis of $y_{new} - \hat{y}_{new}$ [C5, E11] as develop in the SPEC or the ADHER project.

When the number of observations exceeds the number of variables ($n > p$), and when the variables are not correlated, the covariance matrix X^tX is of full rank p. The solution is, in this case, unique and well known: $\hat{\beta} = (X^tX)^{-1}X^tY$.
The High dimensional linear model:
When the number of variables is large compared to the number of observations ($p > n$), or when strong linear dependencies exist, the covariance matrix X^tX is non invertible. As it is technically possible to compute a particularly solution of $\hat{\beta}$ with the help of the pseudo-inverse, this method does not provide any feedback on the informative variables, because of the existence of an infinity of solution. In the "p>n" setup, it is however often the case that a small number of variables brings a substantial explanatory power. Such models which a small number of variables are more interpretable and often preferred. To achieve an accurate estimation, one needs to select the "right" variables and so to determine which parameters $\beta_j, j = 1, \ldots, p$ are not equal to zero. A first approach is to introduce some constraints on the number of coefficients and to compute $\hat{\beta}$ by solving:

$$\min_{\beta} \frac{1}{n}\|Y - X\beta\|_{\ell_2}^2 \text{ subject to } \|\beta\|_{\ell_0} \leq s \tag{1.2}$$

where $\|\beta\|_{\ell_0}$ is the number of nonzero components of β and $s > 0$ a positive parameter. Solving 1.2 is a NP-hard problem and this is not a scalable approach in the high dimensional setup.

1.1 ℓ_1 penalization

The Lasso.
In 1996, the Lasso[1] fundamental paper of Tibshirani brings a first practical answer to the untractable "n>p" setup by replacing the non convex ℓ_0 norm with the ℓ_1 convex norm [Tibshirani, 1996] and proposes to compute $\hat{\beta}$ by solving:

$$\min_{\beta} \frac{1}{n}\|Y - X\beta\|_{\ell_2}^2 \text{ subject to } \|\beta\|_{\ell_1} \leq s \tag{1.3}$$

which is equivalent to minimize the Lagrange form:

$$\min_{\beta} \frac{1}{n}\|Y - X\beta\|_{\ell_2}^2 + \lambda\|\beta\|_{\ell_1} \tag{1.4}$$

where $\lambda \geq 0$ is a regularization parameter. Because of the nature of the ℓ_1 constraint on the coefficients, making the regularization parameter λ value "sufficiently" large induces some of the coefficient values

[1] Least Absolute Shrinkage and Selection Operator

to be theoretically exactly zero and provides a sparse solution [Tibshirani, 1996]. From a computational point of view, in contrast to the ℓ_0 norm, which requires an exhaustive search, the Lasso optimization problem can be solved more efficiently by solving a quadratic programming. When the goal is to select an appropriate value for λ by cross validation, the LARS algorithm introduced by Efron et al. [2004] uses a regularization path, to compute an estimation of β_λ for many values of λ, with a computational complexity of the order of $O(npmin(n,p))$.

The Dantzig selector.
In 2005, the Dantzig selector introduced by Candes and Tao [Candes and Tao, 2005], proposed to solve the ℓ_1 minimization on the coefficients with a regularization on the residuals:

$$\min \|\beta\|_{\ell_1} \text{ subject to } |\frac{1}{n}X^T(Y - X\beta)|_\infty \leq \lambda \tag{1.5}$$

Both procedures Lasso and Dantzig are computationally efficient and adapted to the high dimensional setup. The Lasso procedure can be achieved using a quadratic program and the Dantzig procedure using a linear program. Either from a theoretical or from a practical point of view, they exhibit similar behavior [Bertin et al., 2011, Bickel et al., 2009b, Efron et al., 2007].

Key assumptions.
To guarantee nice statistical properties for the Lasso or the Dantzig procedures, key assumptions on the sparsity of β and on properties of the Gram matrix are needed. Different types of indices have been introduced these last years to provide assumptions on the Gram matrix, $G = \frac{1}{n}X^tX$, and to prove oracle inequalities. A detailed overview of all these restrictive assumptions for prediction and selection is available in [Van De Geer et al., 2009]. Massart et al. [2011] analyze the performance of the LASSO as a regularization algorithm rather than a variable selection procedure.

Candes and Tao [2005] introduced the first indices, the *S-Restricted Isometry Constant*, δ_S^X, and the *Restricted Orthogonality Constants*, $\theta_{S,S'}$, to characterize a design matrix X for sparse recovery purpose:

- δ_S^X, is defined by the smallest number such that $(1 - \delta_S^X)\|\beta\|_2^2 \leq \|X\beta\|_2^2 \leq (1 + \delta_S^X)\|\beta\|_2^2$ for every vector $\beta \in \mathbb{R}^p$ with $\|\beta\|_{\ell_0} \leq S$.

- $\theta_{S,S'}$ computes the restricted correlations between two sparse vectors $\alpha, \beta \in \mathbb{R}^p$ with disjoint sets and is defined as the smallest quantity such that $| < X\beta, X\alpha > | \leq \theta_{S,S'}\|\beta\|_{\ell_2}\|\alpha\|_{\ell_2}$ with $\|\alpha\|_{\ell_0} \leq S$ and $\|\beta\|_{\ell_0} \leq S'$ Small values of restricted orthogonality constant indicate that disjoint subsets of covariates span nearly orthogonal subspaces.

Theoretical results.
For the Dantzig selector, Candes and Tao [2007] first established oracle inequalities in selection, *Under Uniform Uncertainty Principle (UUP)*, which roughly said that for any small set of predictors, the S vectors are nearly orthogonal to each other, and that the model is identifiable.

Theorem 1 (Candes and Tao [2007]). *Considering X a (n,p) design matrix, for any S sparse vector, $\beta \in \mathbb{R}^p$, such that $(\delta_{2S} + \theta_{S,2S} < 1)$, when choosing $\lambda = \sqrt{2\sigma^2 \log(p)/n}$ then with large probability, the estimation $\hat{\beta}$ computed with the Dantzig selector obeys:*

$$\|\beta - \hat{\beta}\|_{\ell 2}^2 \leq \square S\sigma^2 \frac{\log p}{n}$$

where $\square$ represents a universal constant.

Donoho [2006], Meinshausen and Yu [2009] introduced another index, called the S-sparse minimal eigenvalue of G, $\varphi_{\min}(S)$, to characterize the sparsity of the design:

$\varphi_{\min}(S) = \min_{\beta:||\beta||_{\ell_0} \leq \lfloor S \rfloor} \frac{\beta^T G \beta}{\beta^T \beta}$.

Restricted eigenvalues assumption leads also to similar oracle inequalities (but with different universal constants) for the Lasso or the Dantzig procedure in selection (similar to theorem 1) or in prediction (theorem 2), [Bickel et al., 2009b, Meinshausen and Yu, 2009].

Theorem 2 (Bickel et al. [2009b]). *X is a (n, p) design matrix. For any S-sparse vector of parameters, $\beta \in \mathbb{R}^p$, such that $\varphi_{\min}(2S) > c\theta_{S,2S}$, c constant, when choosing $\lambda = A\sqrt{\sigma^2 \log(p)/n}$, $A > 2\sqrt{2}$, then with large probability, the prediction $X\hat{\beta}$ computed with the Dantzig selector or the Lasso, obeys:*

$$||X\beta - X\hat{\beta}||_{\ell 2}^2 \leq \square S\sigma^2 \log p$$

where $\square$ represents a universal constant.

The relations between all previous indices are discussed in Bühlmann and Van De Geer [2011].

From a practical point of view, the restricted isometry, orthogonality, or eigenvalue indices are unrealistic to verify for a given design matrix X when p is large and when the number of coefficients S is not too small. Computing the S restricted isometry constants δ_S of the design matrix for all subset variables is equaled to $\begin{pmatrix} S \\ p \end{pmatrix}$. However, it should be mentioned that the bounds of those indices are theoretically known for some specific matrices as Gaussian ensembles [Candes and Tao, 2005]. The Gaussian assumption is usually not verified in real life applications and checking those conditions is (currently) intractable.

Candès and Plan [2009] introduced the empirical *coherence* τ_n to prove oracle inequalities for the Lasso procedure, defined by:

$$\tau_n = \sup_{\ell \neq m} \frac{| < X_{.,\ell}, X_{.,m} > |}{||X_{.,\ell}||_{\ell_2} ||X_{.,m}||_{\ell_2}}$$

A condition on the coherence, $\tau_n \leq \square / \log(p)$, associated with a sparsity assumption on β, $\beta \in \mathbb{R}^p$ ($||\beta||_{\ell_0}|| \leq S$) and $S \leq \square p / ||X||^2 \log(p))$ shows that the Lasso estimate with $\lambda = 2\sqrt{2\log(p)/n}$ has, with a large probability, a prediction ℓ_2 error similar to theorem 2.

At the opposite to other indices, the empirical coherence can be, in practice, easily computed to check the theoretical assumptions.

Theorem 3 (Candès and Plan [2009]). *Considering X an (n, p) design matrix with $\tau_n \leq c/\log(p)$ (c constant) and the parameter $\beta \in \mathbb{R}^p$ taken form the generic S sparse model, ($||\beta||_{\ell_0}|| \leq S$), such that $S \leq c_0 p/[||X||^2 \log(p)]$, ($c_0$ constant) the Lasso estimate with $\lambda = 2\sqrt{2\log(p)/n}$, with a large probability, satisfies*

$$||X\beta - X\hat{\beta}||_{\ell 2}^2 \leq \square S\sigma^2 \log p$$

where $\square$ represents a universal constant.

Two-step procedures

Using the Dantzig or the Lasso procedure with a regularization factor defined as in Theorem 1 leads to bias in estimating the sparse regression coefficients [Candes and Tao, 2007]. To reduce the bias, two-step procedures involving ℓ_1 penalization, have been introduced in the HD regression framework. The penalized methods are first used to select a first bunch of co variables, then the final estimated coefficients are computed by OLS on the pre-selected variables [Candes and Tao, 2007, Candès and Plan, 2009].

Fan and Lv [2008] provide an alternative two-step procedure called Sure Independence Screening (SIS) for a HD linear model. The SIS procedure selects the "first most" correlated covariates with the target variable Y (in absolute value), then the coefficients of the restricted model are estimated using the Dantzig Selector or the Lasso. Using an intensive simulation, they showed the benefits of a first selection of variables before using the Lasso.

The Dantzig or the Lasso procedures both rely on optimization. In ultra-large dimensions, the computational optimization cost of these procedures is a potential hurdle ant it is therefore interesting to study procedures well designed for ultra-high dimensional models, associated with easy checkable conditions on the data.

1.2 Thresholdings

In a joint work with D. Picard and K. Tribouley, we developed an alternative procedure for the linear model in the "p>n" setup, without any optimization phase and with easily checkable assumptions. The essential motivation was to provide a very simple procedure, based on *a small number of thresholding steps* easy to use in practice and with good theoretical properties. This procedure, called Learning Out of Leaders (LOL), can be viewed as a simple "explanation" of ℓ_1-minimizations [A8].

1.2.1 Learning out of Leaders

In the LOL procedure, the design matrix X is assumed to have normalized columns such that:

$$\frac{1}{n}\sum_{i=1}^{n} X_{i\ell}^2 = 1, \quad \forall\, \ell = 1\ldots, p. \tag{1.6}$$

For clarity, the procedure is here presented with the help of a pseudo-code. As input, the LOL procedure requires data for the target Y, the design matrix X, and the value of two tuning parameters λ_1 and λ_2 which define the level of the thresholds. The outputs of the procedure are the estimated coefficient $\hat{\beta}$ and the predicted target $\hat{Y} = X\hat{\beta}$.

$$(\hat{Y}, \hat{\beta}) \leftarrow \mathrm{LOL}(X, Y, \lambda_1, \lambda_2)$$

Table 1.1: Definition of LOL procedure: input= $(X, Y, \lambda_1, \lambda_2)$, output= $(\hat{\beta}, \hat{Y})$

• **Initialization.** An upper bound on the number of predictors, N^*, that may be selected during the procedure is computed. This bound depends on the empirical coherence τ_n and on a precision parameter ν (Table 1.2):

$\nu = 0.5$ for example	$\nu \in]0,1[$
$\tau_n \leftarrow n^{-1} \max_{\ell \neq m} \lvert \sum_{i=1}^{n} X_{i\ell} X_{im} \rvert$	empirical coherence
$N^* \leftarrow \lfloor \frac{\nu}{\tau_n} \rfloor$	upper bound for the cardinal of the leader set

Table 1.2: LOL initialization part.

The theoretical performances of the LOL procedure are driven by the empirical coherence as will be presented in theorem 4. In practical applications, the value of this index is used to check the ability of LOL to perform the regression, before any computation.

• **Step1: thresholding.** The LOL procedure solves the problem of the choice of the regressors in a crude way by adaptively selecting the regressors which are the "most correlated" with the target and which shown an absolute correlation higher than the λ_1 threshold. The selected regressors are called "the leaders".

For $\ell = 1 : p$	
$\quad \widetilde{\beta_\ell} \leftarrow \frac{1}{n} \sum_{i=1}^{n} X_{i\ell} Y_i$	Compute the 'correlations'
$\quad \widetilde{\beta_\ell}^* \leftarrow \widetilde{\beta_\ell} \mathbb{I}\{\lvert \widetilde{\beta_\ell} \rvert \geq \lambda_1\}$	Threshold
EndFor	
$\mathcal{B} \leftarrow \{\ell, \widetilde{\beta_\ell}^* \neq 0\}$	Set of leaders
If $\#\mathcal{B} > N^*$	
$\quad$ indices $\leftarrow$ order$(\lvert \widetilde{\beta} \rvert)$	Order the largest 'correlations'
$\quad \mathcal{B} \leftarrow$ indices$[1 : N^*]$	Take the indices associated to the $N-$th largest
End(if)	

Table 1.3: "Find the leaders" (LOL/step1)

• **Step2: OLS.** As the number of leaders is then lower than N^* (and not correlated for small empirical coherence), the regression is now stable, and LOL procedure regresses the target on the leaders (Table 1.4):

$\widehat{\beta}_{\lvert \mathcal{B}} \leftarrow (X_{\lvert \mathcal{B}}^t X_{\lvert \mathcal{B}})^{-1} \Phi_{\lvert \mathcal{B}} Y \quad$ {Least square estimators}

Table 1.4: Ordinary Least Square on the leaders (LOL/step2)

• **Step3: thresholding.** The LOL procedure thresholds again the estimated coefficients taking into account, at this step, the noise level (Table 1.5).

The second step selects then the most significant coefficients given the noise of the model. When the value of the empirical coherence is weak (close to zero), $\tau_n \sim 0$ and when the covariables are normalized 1.2.1, it should be noted that the correlations computed, at step 1, provide a direct estimation of the coefficients: $\widehat{\beta} \sim X^t Y / n$.

$\widehat{\beta}^*_{\mid\mathcal{B}} \leftarrow \widehat{\beta}_{\mid\mathcal{B}}\, \mathbb{I}\{\lvert\widehat{\beta}_{\mid\mathcal{B}}\rvert \geq \lambda_2\}$	Threshold
$\widehat{\beta}^*_{\mid\mathcal{B}^c} \leftarrow 0$	

Table 1.5: Last thresholding on the estimated coefficients (LOL/step3)

Illustration of the Learning Out of Leaders procedure

The following examples outline the simplicity of the LOL procedure. We consider the classical framework where the predictors are realizations of Gaussian variables. Observations are simulated from the model $Y = X\beta + \varepsilon$, ε is a Gaussian vector with a Signal over Noise Ratio equaled to 5 (SNR $= 5$). To facilitate the model interpretation, we take S nonzero coefficients equaled to $\beta_l = 2$. A high sparsity example ($S = 10$, left figures) and a lower sparsity example ($S = 50$, right figures) are analyzed for a HD framework with $n = 400$ and $p = 2000$ ($\tau_n \sim 0.25$).

Figure 1.1 illustrates the first step of LOL for the selection of the leaders. All the scalar products $|X_\ell{}^t Y|$ for $\ell = 1, \dots p = 2000$ are computed and represented on the graph; the threshold λ_1 is indicated with a horizontal line (the data driven calibration of the threshold will be discussed later). The leaders, above the λ_1 threshold, are labeled with a small blue cross. The variables which should be rightly selected are indicated with a green dot (true model). When the sparsity is high (S small as in the left part of figure 1.1) and when the predictors are independent, the values of the scalar products of the predictors, really involved in the model, are close to the value of the coefficients $|\beta_\ell| \sim 2$.

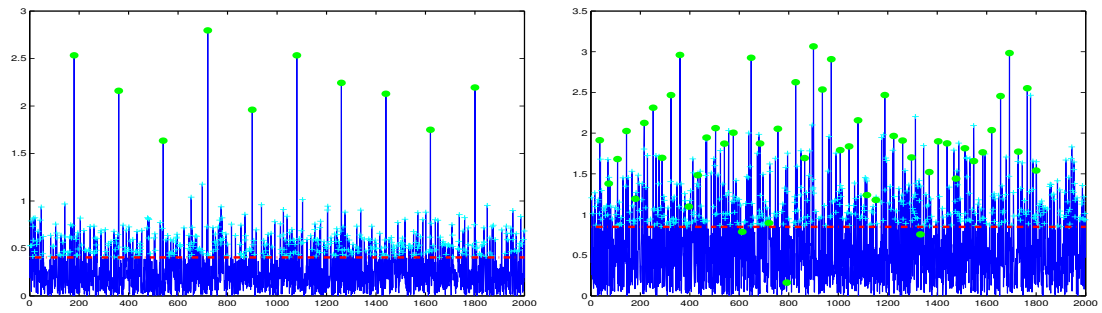

Figure 1.1: $< X_\ell, Y >$, $\ell = 1...p$, $p = 2000$, $n = 400$. Left: $S = 10$; right: $S = 50$. The horizontal line is the first auto-driven threshold, λ_1.

We observe that, in Figure 1.1 left, the reduction of the dimension is very high, after the first step of the procedure: $N = 144$ leaders over the $p = 2000$ initial predictors are selected. In this case, the variables, associated with a (real) nonzero coefficient, are all selected in the leader set. In the right part of Figure 1.1, the sparsity decreases to $S = 50$, some scalar products associated to theoretical significant coefficients fall under the threshold λ_1, the corresponding variables are not selected as leaders during the first thresholding step. In this case, three variables, which should be kept, are eliminated during the first step and are definitively lost for selection.

Figure 1.2 illustrates the effect of the second thresholding on the estimated coefficients after OLS. For the $S = 10$ experiment, the set of zero and nonzero coefficients $\widehat{\beta_\ell}$ are well separated. The procedure performs in this situation quite well neither false positive (FP) nor false negative (FN) are observed. For the right experiment, the separation between both clusters is not so straight and misses detections

(triangular pattern) as false detections (cross not circled) are observed and in this particularly case FN= 8 and FP= 7.

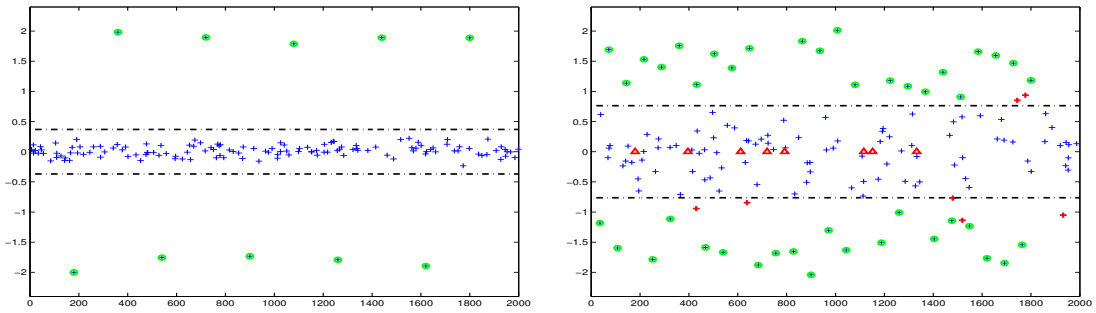

Figure 1.2: Estimations of β_ℓ coefficients with $p = 2000$, $n = 400$. Left: $S = 10$. Right: $S = 50$. The horizontal line is the last auto-driven threshold, λ_2 (see calibration section).

1.2.2 Main theoretical results

It is well known that, when the regressors are normalized and orthogonal, ℓ_1-minimization corresponds to soft thresholding which itself is close to hard thresholding [Bühlmann and Van De Geer, 2011]. In this case, it is natural to expect that thresholding should perform well, at least, in cases that are not too far from the orthonormal conditions which correspond to small coherence conditions.

The following sections present the theoretical performances of the LOL procedure in prediction and in selection. Two different viewpoints are presented: when the empirical coherence is supposed to be upper bounded by $\sqrt{\log(p)/n}$ or not.

For prediction, the criterion of performance, denoted by $d(\hat{\beta}^*, \beta)^2$ is defined as the empirical quadratic distance between the predicted variables and their expected values:

$$d(\hat{\beta}^*, \beta)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{Y}_i - \mathbb{E}Y_i\right)^2$$

case 1: Sparsity assumption and upper bounded coherence

The sparsity constraint on β is introduced using the sets $B_0(S, M)$ (or $B_q(M)$), $M > 0$, [A8] where:

$$B_q(M) := \{\beta \in \mathbb{R}^p,\ \|\beta\|_{l^q} \leq M\} \quad \text{for} \quad q \in (0, 1],$$

and

$$B_0(S, M) := \{\beta \in \mathbb{R}^p,\ \sum_{j=1}^{p} I\{\beta_j \neq 0\} \leq S,\ \|\beta\|_{l_1} \leq M\}.$$

For simplicity, theoretical results are stated only with the $B_0(S, M)$ constraints on the coefficients. Complementary results using l_q balls constraints may be available in [A8].

When the empirical coherence is not too high (upper bounded), the LOL procedure is able to provide a consistent estimation of $Y = X\beta + \varepsilon$, by choosing the adequate thresholds λ_1 and λ_2, even in ultra high dimension, when the ε vector is assumed to be independent Gaussian variables $\mathcal{N}(0, \sigma^2)$ (or sub gaussian). The following exponential bounds are obtained [A8]:

Theorem 4. *Assuming that the coherence* $\tau_n \leq c\sqrt{\frac{\log p}{n}}$, *the number of covariables* $p \leq \exp(c'n)$, *and choosing the thresholds* $\lambda_1 = C_1\sqrt{\frac{\log p}{n}}$ *and* $\lambda_2 = C_2\sqrt{\frac{\log p}{n}}$, *we get for any* $S < \nu/\tau_n$:

$$\sup_{\beta \in B_0(S,M)} \mathbb{P}\left(d(\hat{\beta}^*, \beta) > \eta\right) \leq \begin{cases} 4e^{-\gamma n \eta^2} & \text{for} \quad \eta^2 \geq D\,\frac{S\log p}{n} \\ 1 & \text{for} \quad \eta^2 \leq D\,\frac{S\log p}{n} \end{cases}$$

where ν *is a fixed parameter and* c, c', C_1, C_2, D, η *are positive constants.*

The "naive" approach, proposed by the LOL procedure, requires to have a strong assumption on the value of the empirical coherence ($\tau_n \leq O(\sqrt{\log(p)/n})$) in contrast to many ℓ_1 based methods which only require to have the coherence upper bounded by some constant or by the inverse log-dimension ($\tau_n \leq \frac{c}{\log(p)}$, c constant) as previously presented. The theoretical framework of LOL implies to be very "close" to an identity correlation matrix, in which case the predictors are decorrelated and standard thresholding methods for orthonormal bases may be applied.

Corollary 1. *Under the same assumptions as in Theorem 4, we have:*

$$\sup_{B_0(S,M)} \mathbb{E}d(\hat{\beta}^*, \beta)^2 \leq \square\left(\frac{S\log p}{n}\right)$$

.

• For variable selection, LOL has also the ability to recover properly the regression coefficients:

Theorem 5. *Under the same assumptions as in Theorem 4, we have*

$$\sup_{B_0(S,M)} \mathbb{E}\|\hat{\beta}^* - \beta\|_{l_2}^2 \leq \square\left(\frac{S\log p}{n}\right)$$

.

The convergence rates for the ℓ_2 prediction or estimation error are similar to those obtained for ℓ_1-penalization method (theorems 1, 2).

Under the coherence assumption: $\tau_n \leq c\sqrt{\log(p)/n}$, essential to ensure the RIP conditions, the LOL procedure is minimax [A8].

Focus on the first step:

To emphasize the role of the first step of LOL (in comparison to SIS for instance), we give results concerning the LOL procedure deprived from its second step by enforcing $\lambda_2 = 0$.

Theorem 6. *Assume that* $p \leq O\left(\sqrt{\frac{n}{\log n}}\right)$ *and* $\tau_n \leq c\sqrt{\frac{\log n}{n}}$, *choosing the thresholds* $\lambda_1 = C_1\sqrt{\frac{\log n}{n}}$ *and* $\lambda_2 = 0$, *such that, we get for any* $S < \nu/\tau_n$:

$$\sup_{B_0(S,M)} \mathbb{E}d(\hat{\beta}^*, \beta)^2 \leq \square\left(\frac{S\log n}{n}\right)$$

.

where ν *in* $(0,1)$ *is fixed and* c, C_1, D *are positive constant.*

When the number of regressors p is small, theorem 6 shows that LOL deprived from the second thresholding remains optimal.

case 2: Sparsity and threshold assumption

This section describes another perspective on the performance of LOL. We do not assume anymore that the empirical coherence is upper bounded and satisfies $\tau_n \leq c\sqrt{\log(p)/n}$ and the values of the thresholds λ_1 and λ_2 are known.

The assumptions concern the set $V(S, M)$ of parameters β defined by :

1. *The ℓ_1 norm of the coefficients is bounded by a positive parameter* M: $\|\beta\|_{l^1} \leq M$,
2. *The Sparsity* is defined by the small number of "significant" coefficients, S such that:
   $\#\{\ell \in \{1, \dots, p\},\ |\beta_\ell| \geq \lambda_2/2\} \leq S$
3. which does not exceed the maximum number of leaders, N^*, selected in the algorithm:
   $\sum_{(\ell)>N^*} |\beta_{(\ell)}| \leq c_1 \left(\frac{S \log p}{n\tau_n}\right)^{1/2}$
4. *The biais of leader selection does not exceed the target rate:*
   $\sum_{\ell=1}^{p} |\beta_\ell|^2\ I\{|\beta_\ell| \leq 2\lambda_1\} \leq c_2^2\ \frac{S \log p}{n}$

The right exponential decreasing of the confidence is here achieved on a set $V(S, M)$ of β parameters:

Theorem 7. *Let $S, M > 0$, fix the precision ν in $(0, 1)$. If $\lambda_1 \geq \left(T_1 \left(\frac{\log p}{n}\right)^{1/2} \vee T_2\, \tau_n\right)$ and $\lambda_2 \leq \lambda_1$ (T_1, T_2 positive constants) then:*

$$\sup_{\alpha \in V(S,M)} \mathbb{P}\left(d(\hat{\beta}^*, \beta) > \eta\right) \leq \begin{cases} 4e^{-\gamma n \eta^2} & \text{for} \quad \eta^2 \geq D\ \left(\frac{S \log p}{n} \vee\ S\tau_n^2\right), \\ 1 & \text{for} \quad \eta^2 \leq D\ \left(\frac{S \log p}{n} \vee S\tau_n^2\right) \end{cases} \tag{1.7}$$

D and γ positive constants depending on $\nu, \sigma^2, M, c_0, c_1, c_2$.

The following corollary details the behavior of the expectation of the average prediction error for the LOL procedure, in this case.

Corollary 2. *Under the same assumptions as in Theorem 7, we get*

$$\sup_{V(S,M)} \mathbb{E}d(\hat{\beta}^*, \beta)^2 \leq \square \left(\frac{S \log p}{n} \vee S\tau_n^2\right)$$

for some positive constant D' depending on $\nu, \sigma^2, M, c_0, c_1, c_2$ and r.

Theorem 7 reflects the theoretical behavior of LOL for prediction, even in case of deterioration due to high coherence or a bad choice of the thresholds. The value of the empirical coherence, which can be easily computed on the data, warns the user, before any computation, of the quality of the result. This is a tremendous benefit compared to other assumptions often used for ℓ_1 penalization methods. For a very low empirical coherence value, the complexity of LOL procedure is smaller than most optimization methods.

Comparison with other two-step procedures

The LOL procedure can be connected to the family of orthonormal matching pursuit algorithms as well as to the greedy algorithms [Needell and Tropp, 2009, Tropp and Gilbert, 2007]. The main advantage of LOL compared, with this kind of algorithms, is that there is no iterative search for the leaders. All the leaders are selected in one shot and the procedure stops just after the second step.

Comparing LOL with other two-step procedures as for example, SIS[2]-Dantzig, SIS-Lasso, SIS-SCAD[3], shows that the second step of LOL is much less sophisticated since it consists of computing a least square estimate followed by thresholding. In SIS, the purpose of the first step is essentially to reduce the number of variables, to stabilize and improve the performances of the subsequent optimization algorithms (Lasso, Dantzig or SCAD), so it is truly a preprocessing, as it does not affect the theoretical results since the Dantzig, Lasso or SCAD are already optimal procedures. In SIS, many more regressors are kept since the default value is comparable with $n\log(n)$, whereas the number of leaders in the first step of LOL is bounded by $N = \nu/\tau_n$, which in standard cases is comparable to $\sqrt{n/\log(p)}$. In LOL, the first step affects the results in a much stronger way than for SIS. The second step is not even conceivable without the strong primary variable selection ensuring that thresholding the least square estimate of the coefficients has a meaning (which is not so when it is not uniquely defined).

Of course, as for SIS - giving rise to iterative SIS- LOL can be iterated, and iterative LOL may be also an interesting algorithm. LOL is trading computational complexity for statistical control.

1.2.3 Heuristics for calibration

The (λ_1, λ_2) thresholds are critical parameters for the LOL procedure. The quality of the results depends directly on their values (or their choice). Unfortunately, their theoretical values depend on constant (or parameter as for example σ^2), which are unavailable in practice. At this stage, even if LOL seems to be an appealing procedure from a theoretical point of view, it is perfectly useless if we can not propose a data-driven methodology to calibrate, in practice, those parameters. This is a remaining question which appears for many procedures as for the Lasso or the Dantzig procedures for concerning the choice of the value of the regularization parameter. The performance of the final linear model directly depends on this critical choice. Cross validation is often used to compute adequate data-driven parameters [Arlot et al., 2010, Hastie et al., 2009]. In model selection, Birgé and Massart [2007] proposed the "slope heuristic" data driven procedure to calibrate the penalty [Baudry et al., 2012]. Donoho and Jin [2008] proposed to the high criticism thresholding to achieve optimal phase diagram.

For the LOL procedure, a challenging question is how to choose the threshold based on the data? At step 1 of the procedure, the sparsity and the coherence assumptions suggest that the law of the cross-product (in absolute value) should be a mixture of two distributions: one for the "leaders" (high correlations- positive mean) and one for the others (very small correlations- zero mean). Since the first threshold λ_1 is used to select the leaders, we propose to adaptively split the set of "correlations" $\{K_\ell,\ \ell = 1, \ldots, p\}$, into two clusters in such a way that the leaders are forming one of the two clusters. Many algorithms can be used to split an empirical distribution into two clusters regarding some hypothesis on the underlying data (gaussian mixtures) or not (K means, $K = 2$), hierarchical clustering...

Inspired by the work of Breiman et al. [1984], we propose the following heuristics, based on the deviance function, to compute the thresholds. The boundary between the clusters is computed by minimizing the variance between the two classes after computing the absolute value of the correlations. More precisely, the correlations between each covariate X_j, $1 \leq j \leq p$, and the target Y are computed, $Z_j = |< X_j, Y >|$ then ranked $|Z|_{(1)} \leq |Z|_{(2)} \leq \ldots \leq |Z|_{(p)}$.

[2]SIS: Sure Independence Screening
[3]Smoothly Clipped Absolute Deviation

We consider the deviance function defined by:

$$\mathrm{dev}(J_Z) = \sum_{j=1}^{J} \left(|Z|_{(j)} - \overline{|Z|}^{(J-)} \right)^2 + \sum_{j=J+1}^{m} \left(|Z|_{(j)} - \overline{|Z|}^{(J+)} \right)^2$$

where $\overline{|Z|}^{(J-)}$ and $\overline{|Z|}^{(J+)}$ are the empirical means of the $|Z|_{(j)}$'s for respectively $j = 1, \ldots, J$ and $j = J+1, \ldots, p$. We choose as threshold level

$$\lambda_1 = |Z|_{(\hat{J})} \quad \text{for} \quad \hat{J} = \mathrm{Arg} \min_{j=1,\ldots m} \mathrm{dev}(J_Z).$$

The same procedure is used to threshold adaptively the estimated coefficients $\widehat{\beta}_\ell$ obtained by linear regression on the leaders with $Z = (\hat{\beta}_1, \ldots, \hat{\beta}_m)$ where m designs the number of selected leaders. Again the distribution of the $\widehat{\beta}_\ell$ provides two clusters: one cluster associated to the largest coefficients (in absolute value) corresponding to the nonzero coefficients and one cluster composed of coefficients close to zero, which should not be involved in the model. The frontier between the two clusters, which defines λ_2, is again computed by minimizing the deviance between the two classes of regression coefficients.

This updating procedure is denoted LOLA for "LOL with adaptation". This calibration method of the thresholds was used in all numerical experiments and applications using LOLA algorithm and appears to bring practically good thresholding values in the high dimensional framework [A8, A11, A12, A13]. In practical applications, an additional algorithmic improvement performs a second regression using the final set of selected predictors only involved in the model: the estimators of the (nonzero) coefficients are then slightly more accurate [A8].

1.3 Numerical experiments

When a new statistical procedure is introduced, as the LOL procedure, the practical behavior of the procedure as the impact of the calibration of the parameters (here the data driven thresholds) needs to be fully understood. To fulfill this objective, the following experimental design may be advised:

- Before using the procedure for real applications, it is necessary to evaluate its performances in a practical framework where theoretical assumptions may be checked. In this case, random simulations of the underlying model are particularly well adapted to evaluate the performances for selection and estimation for high to ultra high dimension.
- Investigating practical behavior of the procedure when low coherence assumptions are not strictly satisfied is also particularly informative. This is true for the HD framework when the theoretical assumptions appeared to be somehow strong compared to some practical results. Observing good practical behavior of a procedure, even when going beyond the initial theoretical assumptions may help to relax afterwards the former assumptions in the theoretical framework.
- Challenging to similar "off the shelves" solutions is valuable to evaluate if one procedure may outperform the others in some area.

Before being used for real applications [A11, A13], the LOL procedure has been extensively studied trough simulations [A8, A12]. We briefly recall the main lessons from the experimental design used for the LOL procedure. More details can be found in [A8, A12].

Experimental protocol

If not specified, the $(n \times p)$ design matrix X is filled with i.i.d gaussian variable such that $Y = X\beta + \varepsilon$, β and the signal over noise ratio is given. For such design matrix, the empirical coherence value may reach large value: for example: $\tau_n \sim 0.5$ for $p = 1000$ and $n = 10$. As it is explained before, the value of τ_n provides information, before any computation, about the theoretical accuracy of the LOL procedure. Figure 1.3 (right) shows the evolution of the empirical coherence as a function of $\sqrt{\log(p)/n}$ which may allow to compute the constant c introduced in Theorem 4. We observe that, for a given number of variables, the empirical coherence strongly increases when the number of observations decreases.

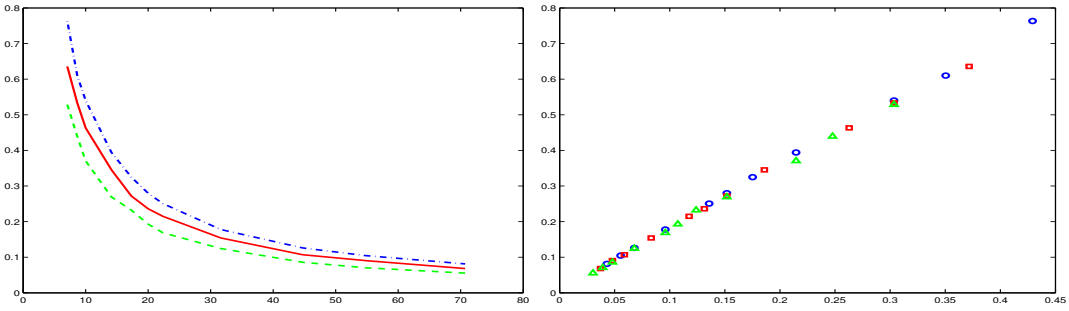

Figure 1.3: $Y-$axis: Average empirical coherence, τ_n, computed from $K = 500$ realizations of a design matrix filled with iid Gaussian variables. $X-$axis: $\sqrt{n}$ (left) or $\sqrt{\frac{\log(p)}{n}}$ (right) for $p = 100$ (dashdot line or triangle -green), $p = 1000$ (solid line or square -red), $p = 10000$ (dash line or circle -blue). $K = 500$

1.3.1 Ultrahigh dimension

The low complexity of the LOL procedure shows particulary benefits in ultra high dimension, for example up to $p = 20000$ for $n = 400$ observations. For small sparsity levels, the performances of LOL reach the inverse of the simulated Signal over Noise Ratio. As expected, the performances decrease when the values of the sparsity level S/n or the indeterminacy $1 - n/p$ increases [A8].

1.3.2 Two-step procedures

Compared to other two-step procedures (SIS-Lasso, SIS-Scad, Lasso-Reg), the LOL procedure shows also good behavior. As expected, all the procedure performances depend on the sparsity and on the noise (SNR). As expected, then LOL procedure shows particularly benefits over the other procedures when the SNR is small or when the sparsity S is high [A8].

1.3.3 Beyond theoretical assumptions

Embedding real data in a high dimensional framework:
The performances of LOL were evaluated using real data but embedded in a high dimensional framework. The data are "the Communities and Crime data" *UCI machine learning data base repository*. The

target variable Y ($n = 1000$ observations) denotes the total number of violent crimes per 100K population and the initial $p = 101$ regressors included in the data set involve indicators of the community, such as the percentage of the population considered as urban, the median family income or involve law enforcement, such as the per capita number of police officers, percentage of officers assigned to drug units... In order to evaluate the LOL procedure, we first select a benchmark model computed using a stepwise procedure where it appears that $p_0 = 14$ regressors are finally selected among the $p = 101$ initial regressors.

The LOL procedure is applied on the previous data, embedded in a high dimensional space by adding artificially variables whose laws mimic the different underlying statistical distributions of the $p_0 = 14$ original variables [A12]. Here, we added 1000 independent random variables distributed according to seven different laws: Normal, lognormal, Bernoulli, Uniform, exponential with scale parameter 2, Student $T(2)$, $T(3)$ in equal proportion. The final size of the regressors set is then $p = 7101$. The results (obtained with $K = 1000$ runs) in terms of predictive error are particularly satisfactory: the prediction error computed for LOL procedure (working artificially in high dimension) is similar to the prediction error computed with the stepwise model (working in low dimension). For the LOL procedure, the error is $0.3519\ (0.0070)$ to be compared to $0.3433\ (0)$ for the OLS methods in small dimensionality $p_0 = 14$. LOL reduces drastically the dimensionality since the selected models are of size $10.2320\ (1.4854)$. The selected variables are among the good ones: less than one artificial variable per run (among 7000 candidates) is wrongly selected [A12].

1.4 Conclusions

The LOL procedure brings a very simple answer to the estimation of the coefficients of a linear model in high dimension. The theoretical behavior of the procedure depends on assumptions on the design matrix (via the empirical coherence), on the underlying linear model (via the sparsity of the coefficients, the noise) and on the choice of two thresholds which are successively applied on the set of cross products between the predictors and the target, and on the estimated coefficients. Associated with the LOL procedure, we propose a simple heuristic, with no cross-validation, which appears to bring appropriate calibration values for various applications. However, it should be underlined that the low complexity of the algorithm has a cost: a strong assumption of the bound of the empirical coherence.

When the low coherence assumption is not satisfied, the theoretical behavior of the procedure is not guaranteed anymore. However, some numerical experiments demonstrated that the practical uses may still behave well. This appropriate behavior has an answer: the empirical coherence depends on the set of correlations computed between all the normalized predictors, and all correlations do not have the same "practical" impact in the LOL procedure. Predictors which show strong correlations and which are not correlated to the target -and so not involved in the true model - will be easily removed at the first step of the procedure, and consequently will not have any impact on the estimation procedure. On the opposite, predictors which are both correlated to the target and correlated to some others predictors are selected as leaders, in the first step. In this case, they induce an instability of the estimation. Introducing a structure of groups in the linear model answers to the question of dealing with intra and inter correlations and is presented later in Chapter 3.

Chapter 2

Functional regression, sparse model and forecast: application to electrical consumption

*T*he work presented in this chapter has been performed through a collaboration with V. Lefieux, L. Maillard-Teyssier from RTE and D. Picard, K. Tribouley from Paris Diderot University. It has been published in:

A13 M. Mougeot, D. Picard, V. Lefieux, L. Maillard-Teyssier. (2015) *Modeling and Stochastic Learning for Forecasting in High Dimension.* Springer Lecture Notes in Statistics, p161-182.

A11 M. Mougeot, D. Picard, K. Tribouley. (2013) *Sparse approximation and fit of intraday load curves in a high dimensional framework.* Advances in Adaptive Data Analysis p 1-23.

—

Nowadays, one big challenge in the industry is to be able to automatically analyze operational data for decision making processes and forecasting is a major issue. RTE, the French electricity transmission system operator, is responsible for operating, maintaining and developing the high and extra high voltage transmission network. RTE should guarantee the supply. As the electrical power cannot be stored, anticipating the French electricity demand is crucial and helps to ensure the permanent balance between generation and consumption at all times. Figure 2.1 shows, as an example, the French National electricity consumption signal during one week of consumption. One major operational need for RTE is to get every day, for example at 5 pm, a consumption forecast for the next day.

During the last years, the electricity consumption models have been continuously upgraded by integrating more and more variables, including calendar effect, demographic and economic variables [Hyndman and Fan, 2010] as well as weather conditions, thanks to the improvement of the sensor technology in climatology (wind, satellite cloud cover, grid temperature...). Models have become complex, hard to analyze, and most of them suffer from an over-parametrization. In a context where electricity consumption strongly evolves, the calibration of all the internal parameters becomes problematic and induces a low adaptability and reactivity of these kind of models.

Our work began in 2011 with a practical question asked from the R&D department of RTE by V. Lefieux and L. Maillard-Teyssier. The question was: *"Is it possible to built forecast models in the electricity*

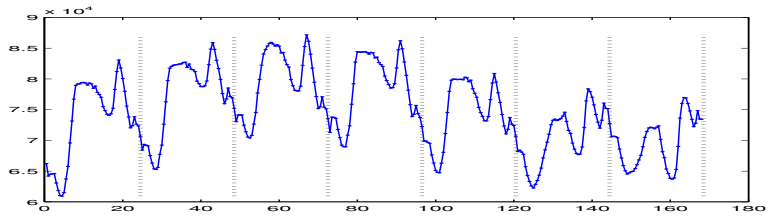

Figure 2.1: one week of electrical consumption signal from Monday January 25th to Sunday January 31th 2010, 30' sample, horizontal units in hours

consumption field which would rely on very few parameters and would be easy to calibrate -without the need of human expertise- and which, at the same time, would show good performances?". At that time, there was no obvious answer, as it was commonly admitted that many contextual variables may be influential for the electricity consumption prediction. However, it was also known that a robust and efficient prediction has to rely on a small number of well chosen predictors. With D. Picard and K. Tribouley, we addressed the RTE questions during two successive academic/industrial collaborations between the LPMA and RTE from 2011 to 2014 [A11, A13]. To answer the question, we investigated the use of HD sparse functional regression, first to model, and then to forecast the intra-day load curves. During this project, we developed a generic method which may be used in other applications requiring forecasting time series which exhibit regularly noisy patterns over periodic time windows.

The main steps of the method are presented hereafter, and will be developed in the following sections.

1. **Data Mining and knowledge discovery of the intra-day load curves.** The first task of the project is dedicated to mine the intra-day load signals on an historical database in order to create relevant inputs for a predictive model.
   - To avoid the effects of dimensionality in the analysis of the set of electrical curves, a first sparse representation of the functional intra-day signals is provided using a generic dictionary of functions.
   - Using the previous sparse representation, the intra-day load signals are studied in a low dimensional space and a clustering analysis is performed to check if some underlying statistical structure may be exhibited in the historical set of signals.
2. **Modeling the intra-day load curve.**
   - Based on the previous cluster analysis, a pattern variable is created, in order to be able to introduce relevant shape information in the predictive model.
   - Besides the pattern variables, exogenous meteorological variables are introduced to built the final predictive model for the intra-day load curves.
3. **Forecasting.** The forecast of the intra-day load curves relies on an information retrieval task. Former estimated coefficients are retrieved and are plugged in the predictive model.
   - The estimated coefficients are chosen given a strategy which relies on a comparison between the calendar or meteorological information of the following day and the same information already available in the historical database (also nearest neighbor method). Different strategies lead to different forecast experts.

- The final forecast is provided using an weighted aggregation of the bunch of statistical experts.

This methodology is today available in a dedicated software delivered to RTE for complementary validation, in a more operational environment [Bourriga and Lefieux, 2014].

2.1 Functional regression and HD linear models

For electrical demand forecasting, various models have already been proposed. Time series analysis have been widely used through various models as for examples ARIMA models [Hagan and Behr, 1987], [Chakhchoukh et al., 2009], nonparametric regression [Poggi, 1994], neural networks Marin et al. [2002] and exponential smoothing [Christiaanse, 1971, Taylor, 2010, 2012]. Aggregation of large sets of time series predictors has also been proposed [Devaine et al., 2013]. Functional data analysis has also been investigated where the daily electricity load is modeled as curves [Antoniadis et al., 2006, Cho et al., 2013, Cugliari, 2011, Devijver, 2014].

In our case, to naturally integrate the strong daily time dependencies observed in the electricity consumption series (Figure 2.1), we model each intra-day signal as a functional data. The entire time series signal Y is split into N sub signals $(Y_1, \dots, Y_t, \dots, Y_N)$ where $Y_t \in \mathbb{R}^n$ denotes the sub signal of length n for the t^{th} day: $Y_t = (Y_{t,1}, \dots, Y_{t,i}, \dots, Y_{t,n})$.

Each day t, the curve Y_t is modeled in a supervised learning setting where each time unit signal, t, is considered as a unknown function f_t to be learned. As the electricity consumption data are regularly spaced (issued every half hour), we observe $Y_{t,i},\ i = 1 \dots n$ with $n = 48$.

The following functional regression model is considered:

$$Y_{t,i} = f_t(i/n) + \varepsilon_{t,i} \quad \text{for } i = 1, \dots, n \tag{2.1}$$

where $\varepsilon_{t,i}$ are iid Gaussian $\mathcal{N}(0, \sigma^2)$ for some positive constant σ^2.

A dictionary made of p functions, $\mathcal{D} = \{g_1, \dots, g_p\}$ is introduced in order to explain the unknown function f_t, which is the written as:

$$f_t = \sum_{\ell=1}^{p} \beta_{t,\ell}\ g_\ell + h_t \tag{2.2}$$

where the g_ℓ functions of the dictionary are normalized with respect to the empirical measure:

$$\forall \ell = 1, \dots, p, \quad \frac{1}{n} \sum_{i=1}^{n} g_\ell^2(i/n) = 1.$$

and h_t is a 'small' function (in absolute value).

Combining (2.1) and (2.2), the functional regression model is written as follows:

$$Y_{t,i} = \sum_{\ell=1}^{p} \beta_{t,\ell} g_\ell(i/n) + h_t(i/n) + \varepsilon_{t,i},\ i = 1, \dots, n$$

and coincides with the linear model:

$$Y_t = X\beta_t + u_t + \varepsilon_t \tag{2.3}$$

where X denotes the matrix with general term $X_{i,\ell} = g_\ell(i/n)$ and where $u_{t,i} = h_t(i/n)$.

In order to estimate f_t, the unknown parameters β_t are computed, by minimizing:

$$\|Y_t - X\beta_t\|_{\ell_2}^2$$

.

In the $p \geq n$ setup, a sparsity condition on β_t should be assumed to solve this problem. The LOLA algorithm, already presented in the previous chapter is used to select and estimate the coefficients $\hat{\beta}_t$:

$$(\hat{Y}_t, \hat{\beta}_t) = \text{LOLA}(X, Y_t)$$

Performances of the fit of each daily curve Y_t is computed using the usual Root Mean Square Error (RMSE) or the Mean Absolute Percentage Error (MAPE) error, more used in the electricity consumption field:

$$\text{RMSE}_t = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_{t,i} - Y_{t,i}\right)^2} \qquad \text{MAPE}_t = 100 * \frac{1}{n}\sum_{i=1}^{n}\left|1 - \frac{\hat{Y}_{t,i}}{Y_{t,i}}\right|$$

Remark. In the functional regression framework, the consistency of the LOL procedure, already presented in theorem 4 of Chapter 1 is guaranteed when the u_i, $1 \leq i \leq n$ are "relatively small" that is when one can make the assumption that $\sup_{i=1,\ldots,n} |u_i| \leq c_0 \sqrt{\frac{1}{n}}$ with c_0 constant. This assumption means that the curve Y_t may be well approximated on the dictionary $\mathcal{D}$ [A8].

2.2 Mining the curves using sparse approximation

One week of electricity consumption signal, as presented in Figure 2.1, exhibits daily patterns. These daily patterns are common when observing times series which correspond to the average of local behaviors of users such as consumption or production. The shape of the day characterizes typical behaviors of electricity consumption. Being able to understand these kind of behaviors on first hand and to link them, for example, to contextual data on a second hand, is a crucial point which may help to reduce the apparent variability of the global problem by introducing contextual analysis. Introducing knowledge of the underlying statistical structure in a model of fit and forecast will improve the model performances. In the field of electricity consumption, it is well known that daily curves are mostly explained by calendar and climate factors. French electricity consumption is known to be larger in winter than in summer and typical profiles can be observed depending on the type of days Cugliari [2011]. In most applications, in order to integrate the calendar information, the set of days is simply split on "apriori" calendar bases. For example, Taylor [2012] uses a partition of size 20. Splitting the different days into 5 groups (Monday, Friday, Saturday, Sunday and the others) subdivided by the four seasons Winter, Spring, Summer and Autumn for French and British consumption data. To study the Spanish consumption, Marin et al. [2002] use Kohonen maps to build groups of consumption. We propose here to learn adaptively the representative 'patterns of consumption' using the sparse representation of the intra-day curves to avoid the effects of dimensionality, on a large set of historical data.

2.2.1 Choice of a generic dictionary

The choice of a generic dictionary, $\mathcal{D}$, to get a sparse representation and a good approximation of the curves is a central issue. At the opposite to the theoretical framework, for 'real' data, we can not just assume that an "adequate" dictionary exists, we have also to built it in practice. A fundamental question is how to choose the functions of the dictionary? Unfortunately, there is no universal answer but it is

however possible to follow some basic guidelines. The nominal shape of the daily signals brings a first answer. Signals exhibiting smooth periodic shapes may be sparsely represented using trigonometric bases. When abrupt variations are observed in a relatively short amount of time, Haar functions may be more appropriate to capture localized irregularities. When the signals exhibit various features as periodicity and abrupt variations, the choice of a mixed dictionary combining different bases may offer a sparser representation than using independently on or the other bases.

For the intra-day load curves, different dictionaries of functions were analyzed. As periodicity and abrupt variations appeared in the signal, Haar, Fourier, Wavelets (DB7) basis and a mixed dictionary composed of Fourier and Haar functions were tested, and the coefficients were computed with the LOLA algorithm. For the French National electricity consumption, $N = 2800$ sub signals of length $n = 48$ (half hour sampling) are extracted from the global consumption signal, sampled every half hour, from January 1^{st}, 2003 to August 31^{th}, 2010. The first practical study shows that the average sparsity, $\bar{S}$ strongly differs between the dictionaries, from a highest sparsity for Fourier basis equaled to $\bar{S}_{trigo} = 5.5$ to a lowest sparsity for DB7 basis equaled $\bar{S}_{DB7} = 9.4$ [A11]. Haar and DB7 dictionaries seem to be the less appropriate bases because they need in average more coefficients, and show the highest RMSE and MAPE errors. The mixed dictionary improves the quality of restoration with respect to the MAPE as well to the RMSE errors. Moreover, the mixed dictionary appears to perform better than the Fourier basis for the approximation of the signals around the peaks of consumption which are well captured with the Haar functions [A11].

Studying the intra-day load signals in a low dimensional space. It appears here that, a subset of 20 functions over the $p = 62$ initial functions of the mixed dictionary are called to adjust 99% of the daily signals [A11]. The set of signals can then be analyzed in a lower dimensional space using their sparse representations, because all the intra-day load signals share the same subset of functions of the dictionary. It should be underlined that unless introducing any constraints for computed the approximation of the intra-day signals, no guarantee exists that the set of curves share the same support.

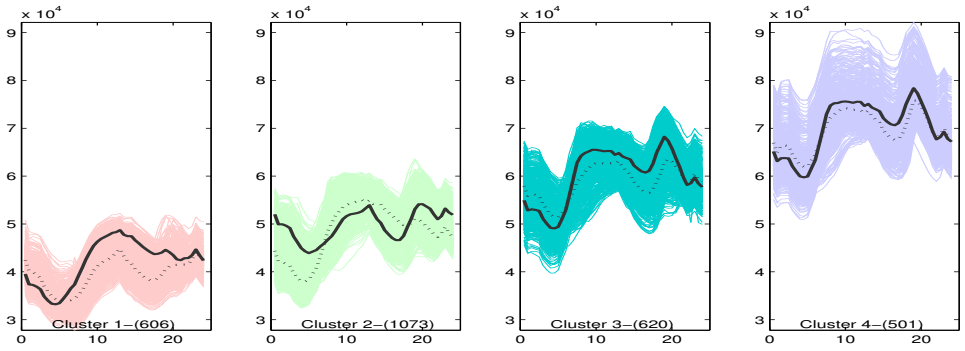

Figure 2.2: Clustering of the 2800 intra-day load curves using a two step K means algorithm: size (4 main clusters in color) and shape (centroids of subclasses in gray) effects.

2.2.2 Mining and clustering to define adaptive patterns of consumption

When receiving data from a given application, a first usual task is to mine the data to get a global understanding and to analyze if any specific underlying structure may be observed which may be helpful for predictive modeling. In marketing applications, for example, this task is systematically done to provide a previous segmentation of consumers before any modeling. In the electricity consumption application, in order to catch potential structural information, a segmentation of the set of the intra-day load curves is performed. To avoid to damage the classification by the curse of dimensionality and to get robust results, it is crucial to use data which do not lie in a too large dimensional space. For that reason, we use the previous sparse representation of the curves on the mixed dictionary to perform the clustering. At this step, many different algorithms may be used for the clustering. We choose the K-means algorithm for its simplicity but associated with an analysis of the stability of the number K, in order to catch the right number of groups [A11]. For the electricity consumption application, $K = 8$ different clusters emerged, characterized with different size and shape as shown in Figure 2.2. At this step, each cluster may be summarized by a *pattern of consumption*. The patterns of consumption are the different curves which correspond to the centroid of each cluster.

2.2.3 Patterns of consumption as endogenous variable

Mining the set of historical curves is mainly a descriptive task and the result of the K-means algorithm i.e. the adaptive clusters is, at this step, useless for predictive modeling as, in particular, the curve Y_t has to be known in advance to be able to get its associate pattern of consumption. In order to exploit the underlying structure of adaptive clusters for predictive modeling, one needs to be able to describe each cluster with the help of generic variables, as for example calendar variables which may be known before prediction (in order to be used in the prediction). This correspondence makes then possible to characterize in advance any day, first by its calendar status then by an associated pattern. In the context of forecast, the calendar interpretation of the clusters is absolutely necessary. For the electricity consumption application, in order to understand the potential calendar features caught by the clusters, the distribution of the type of days and months has been first analyzed for each computed cluster. For the set of intra-day load curves, an interpretation of each cluster in term of calendar statements is provided. Each cluster is described using a code defining a period of the year using the day (1 to 7 from Monday to Sunday) and the month (1 to 12 for January to December); the set of periods making a partition of the year [A11].

Based on the calendar interpretation of the clusters, a *pattern variable* called G_t is defined. G_t has the same number of modalities than the number of clusters as presented in figure 2.3. Each modality of G_t is computed as the average of the intra-day load curves on an historical set of data given the definition of corresponding days and months of the cluster. It sould be noted that due to the calendar reinterpretation, the modalities of G_t may be slightly different to the centroids computed with the K means algorithm.

2.3 Modeling

Up to now, the main objective of this work was to find a generic dictionary to get a sparse representation of the signals, to perform a clustering for mining the approximated curves and to define a pattern variable computed on a calendar representation of the clusters. The following section introduces the model which will be finally used to forecast.

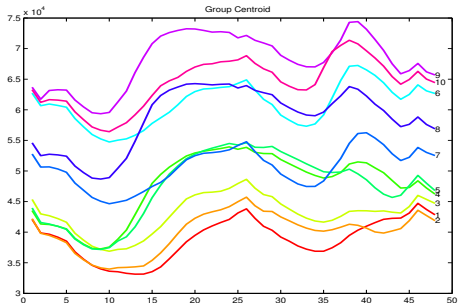

Figure 2.3: Modalities of the pattern variable G_t. In this specific case, 10 modalities are computed (2 more groups for specific "ejp" days and Christmas vacations)

2.3.1 Sparse model and adaptive dictionary for modeling the curves

Based on the previous work of segmentation, we propose to approximate the intra-day signal Y_t using an adaptive dictionary. The design matrix of the generic dictionary X is replaced, in equation 2.3, by $X_t = [P_t\ M_t]$ where:

- $P_t = [G_t\ Y_{t-7}]$ is the concatenation of G_t the group variable previously defined and of Y_{t-7} the intra-day curve one week before t. P_t defined the **pattern variables**. The size of P_t is $(48, 2)$.
- M_t defines a non linear summary of the variations and sizes of temperature, cloud cover or wind all over the French territory [A13]. The size of M_t is $(48, 12)$. M_t defined the **Exogenous variables**.

The linear model for the daily load curve is

$$Y_t = X_t\beta_t + u_t + \varepsilon_t \tag{2.4}$$

where the unknown parameter β_t belongs to $\mathbb{R}^p$ $(p = 14)$.

It should be stressed that the adaptive dictionary, X_t, is directly issue of the first data mining task, and that it is designed to offer a very high sparsity representation. We observe than, in average, $\bar{S} = 2.5$ nonzero coefficients are used to approximate the last year of electricity consumption defined by 365 intra-day load signals from September 1^{st} 2009 to August 31^{th} 2010), with a average MAPE error equaled to 1.24% (median 1.05%). On the same period, the mixed dictionary defined with generic functions provides an average sparsity of $\bar{S}_{mixed} = 7.0$, with an average MAPE error of 1.43% [A11]. In this case, it should be underline that we do not use the benefit of the Learning Out of Learning Algorithm to work in a high dimensional framework but, either its ability to select the most appropriate functions of the dictionary.

2.4 Forecasting

To forecast the intra-day load curve $\widetilde{Y}_t$ of the day t, we refer to the previous linear model and writes:

$$\widetilde{Y}_t = X_t\ \widetilde{\beta}_t$$

The matrix $X_t = [P_t\ M_t]$ is known before t.

- $P_t = [G_t\ Y_{t-7}]$ defines the patterns of day t. Based on calendar statements, the modality of G_t is known before t as the curve Y_{t-7} which is the one week ahead intra-day load curve.

- The meteorological variables, M_t, are here supposed to be known. In real applications, these variables will be provided by Meteo France, the French agency for weather prediction.

The main issue here is to provide $\widetilde{\beta}_t$. Our approach will be to chose a "good candidate" for $\widetilde{\beta}_t$, among the set of already estimated coefficients $\widehat{\beta}_u$ computed for the past intra-day load curves ($u < t$). This strategy is motivated by the fact that the linear model introduced in equation (2.4) appears to be a good model to approximate the intra-day load curve and is moreover a sparse model which relies only on a small number of coefficients.

2.4.1 The experts

For one day, similar causes of weather or calendar conditions or identical groups of consumption should provide similar effects and then a similar electricity consumption. A collection of expert forecasters is here introduced. Each expert has its own strategy which consists to compare the day t at hand to referring scenarios extracted from the past i.e. finding, in the past, a day t^* which is closest according to its strategy to the day t. In order to retrieve t^*, different *strategies* are introduced. A strategy, called s, is a function defined from $\mathcal{T}$ to $\mathcal{T}$, where $\mathcal{T}$ denotes the set of indices of the different days. For any $t \in \mathcal{T}$, we have: $s(t) = t^* < t$. A forecasting *Expert* is then simply associated to a strategy s and provides a forecast of the intra-day load signal of the next day t by plugging-in the approximated coefficients $\widehat{\beta}_{s(t)}$ calculated at day $s(t)$ chosen by strategy s:

$$\widetilde{Y}_t^{\,s} = X_t \widehat{\beta}_{s(t)}$$

A practical question is 'How to choose the experts' ? Many factors are known to have a potential impact on the electricity consumption. Time-lag specialized experts are introduced which simply retrieve the estimated coefficients corresponding to the day before or the day, one week before. Meteorological experts retrieve the estimated coefficients corresponding to the closest temperature, nebulosity or wind using ℓ_2 or sup distances. Up to 17 strategies are introduced, in this application, to potentially forecast the intra-day load curves [A13].

2.4.2 Aggregation

It appears that the experts perform independently well depending on days, or meteorological issues. But no one among them achieves the best performance most all the time [A13]. There is an obvious need to combine them. In the recent years, many interesting theoretical results as well as practical simulations have been obtained using aggregation and especially exponential penalization: see Juditsky and Nemirovski [2000], Catoni [2004], Dalalyan and Tsybakov [2007], Tsybakov [2003]. Among those references, some of them are dedicated to the prediction of time series or individual sequences, Devaine et al. [2013], Gaillard and Goude [2014]. A crucial problem is however to find appropriate weights for each expert. In this context of prediction, this is a challenging issue which can give rise to very sophisticated procedures. For the sake of simplicity we present here a very understandable and manageable one, which only records the approximation properties of each expert and penalizes those with poor approximation results. More precisely, let us recall that $\mathcal{M}$ is the set of strategies introduced above, and $\widehat{Y}^s$ the forecasting expert computed with the strategy s. The aggregated expert is a weighted sum of all the consumption forecasts provided by the different experts:

$$\widehat{Y}_t = \frac{\sum_{s\in\mathcal{M}} w_t^s \widetilde{Y}_t^s}{\sum_{s\in\mathcal{M}} w_t^s}$$

where w_t^s are positive weights depending on the day t and the strategy s.

As explained above, our procedure penalizes by putting small weights, on the strategies which were not able to approximate well the signal at $s(t)$: e.g. the weights w_t^s depend in an exponential way on the l_2 error of $\|Y_{s(t)} - \hat{Y}_{s(t)}\|_2^2$:

$$w_t^s = \exp(-\|Y_{s(t)} - \hat{Y}_{s(t)}\|_2^2/\theta)$$

$\theta > 0$ is a standard tuning parameter (also called temperature parameter with reference to statistical physics). Practically, this parameter is chosen using cross validation on the past. Using aggregation with exponential weights, we observe that the MAPE is much smaller than the different errors computed for each individual experts showing the benefits of the different contributions.

2.4.3 Performances

Figures 2.4 and 2.5 give a graphical illustration of forecast for two different weeks chosen in winter and spring. We observe that forecasts are more accurate during spring periods than winter periods. In Figure 2.5, local maxima seem to be overestimated, while local minima are underestimated.

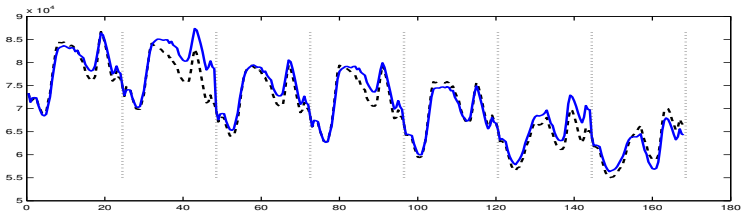

Figure 2.4: Forecast (solid blue line) and observed (dashed dark line) electricity consumption for a winter week from Monday February 1st to Sunday January 7th 2010.

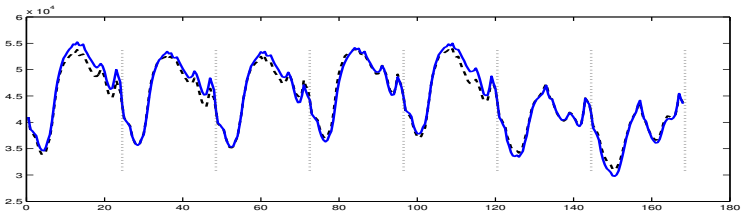

Figure 2.5: Forecast (solid blue line) and observed (dashed dark line) electricity consumption for a spring week from Monday June 14th to Sunday June 21th 2010.

If we were able to find each day the best strategy to apply, the oracle MAPE average error over one year would be equal to 1.44% (standard deviation 0.74%), which is very similar to the approximation MAPE error of 1.70% computed over the same period and a satisfactory performance for these prediction experts. The aggregation of the specialized experts shows an error MAPE of 2.18% (std 1.2%) which competes favorably with actual automatic operational forecasts (before any manually tuning) [A13].

2.5 Software

As previously mentioned, the modules for functional regression, sparse modeling estimation and forecast have been developed during two successive collaborations between RTE and LPMA from 2011 to 2013 [A11], [A13]. All the algorithmic work of the project was delivered to RTE in 2011 for the modeling part and in 2013 for the forecasting part. The Matlab codes are currently running at RTE which investigates the performances of the methodology on new data [Bourriga and Lefieux, 2014].

2.6 Conclusions

Theoretical and practical advances have been required in order to meet RTE's demand. The Learning Out of Leaders algorithm has been applied for the functional regression models. The data mining analysis on the historical set of data which is always a very time consuming task, was essential to built adaptive dictionaries providing sparse models for the intra-day load curves, and to indirectly guarantee the theoretical assumptions needed for LOL procedure.

The methodology based on sparse functional regression models and aggregation of experts appears to be generic, and may be applied to other areas as for example water consumption forecast or monitoring.

As usual, the presentation of this work, at this stage, raises immediately new questions and new perspectives of practical and theoretical works.

- Concerning the reduction of dimension of the set of intra-day load curves, performed before the clustering task, the work on the Grouping Learning Out of Leaders procedure (see Chapter 3) should help to implement multi-task learning to guarantee a reduction of dimension and, at the same time, the same support across all the curves.
- To improve the forecast, more experts will be introduced in the future. The method of aggregation should be diversified according to the feedback of the short term forecasting platform.
- Due to the operational needs, different adaptations of the forecast will be provided. Particularly, the horizon forecast should be extended to 48 hours, or more and the method should be adapted to choose the delivery time of prediction, according to business constraints.
- Finally, confidence intervals should be provided for the forecast, which raise novel theoretical questions.

Chapter 3

Grouping the variables for HD linear models

*T*he work presented in this chapter has been performed through a collaboration with D. Picard and K. Tribouley. It has been published in:

A10 M. Mougeot, D. Picard, K. Tribouley. (2013) *Grouping Strategies and Thresholding for High Dimensional Linear Models rejoinder.* Journal of Statistical Planning and Inference 143, p 1457-1465.

A9 M. Mougeot, D. Picard, K. Tribouley. (2013) *Grouping Strategies and Thresholding for High Dimensional Linear Models, with discussion* Journal of Statistical Planning and Inference 143, p 1417-1438.

—

With the hope of taking advantage of prior knowledge, relations between covariates may be introduced in a linear model. For example, in gene expression analysis, genes from the same biological pathway can be considered as belonging to the same group and, in this situation, it is desirable to take into account, in the analysis of such data, an "a priori" group structure [Huang et al., 2012]. Moreover, such a group structure may improve the prediction performance and/or the interpretability of the models compared to standard HD linear models, [Friedman et al., 2010].

We consider here **the HD linear model** studied in Chapter [1]:

$$Y_i = X_i\beta + \varepsilon_i, \quad i = 1, \dots, n \tag{3.1}$$

with a group structure on the predictors, ε_i is a non observed random Gaussian error, $N(0, \sigma^2)$.

In this Chapter, k denotes the number of predictors and *the set of predictors is distributed into* p *groups* denoted by $\mathcal{G}_1, \dots, \mathcal{G}_p$ with $\mathcal{G} = \mathcal{G}_1 \oplus \dots \oplus \mathcal{G}_p$ defining a *partition* of the k initial predictor indices. A same predictor can not belong to different groups (non-overlapping case) and we consider that the number of predictors k exceeds the number n of observations. An important application of HD linear models with groups is the *multi-task learning*. In this specific case, T successive linear models (the tasks), $1 \leq t \leq T$, are considered:

$$Y_i^{(t)} = X_i^{(t)}\beta^{(t)} + \varepsilon_i^{(t)}, \quad t = 1, \dots, T, \quad i = 1, \dots, n_0 \tag{3.2}$$

with n_0 observations and p common covariates between the tasks. This set of models can be reformulated as a single regression problem with a structure of groups by setting $k = pT$, $n = n_0 T$ and identifying the vector of coefficients $\beta \in \mathbb{R}^p$ by the concatenation of the vectors $\beta^{(1)}, \dots \beta^{(T)}$. The groups $\mathcal{G}_j$, $1 \leq j \leq p$ are defined by the tasks across the T initial linear models.

3.1 Group penalization

Considering HD linear models with a structure of groups, each group of variables is treated as a "unit" [Friedman et al., 2010, Huang et al., 2012]. All predictor coefficients in the same group should behave similarly: simultaneously be estimated or discarded. In 2006, the **Group Lasso**, as a natural extension of the Lasso, introduced in the penalty function the ℓ_2 norm of the coefficients for groups of variables [Yuan and Lin, 2006] to solve the optimization problem:

$$\min_{\beta} \frac{1}{n} \|Y - \sum_{j=1}^{p} X_{\mathcal{G}_j} \beta_{\mathcal{G}_j}\|_{\ell_2}^2 + \sum_{j=1}^{p} \lambda_j \|\beta_{\mathcal{G}_j}\|_{\ell_2} \tag{3.3}$$

where $\lambda_1, \dots, \lambda_p$ are positive regularization parameters (which may be equaled). $X_{\mathcal{G}_j}$ and $\beta_{\mathcal{G}_j}$ correspond respectively to the sub design matrix and the coefficients of the group $\mathcal{G}_j$ ($1 \leq j \leq p$).

Assuming a group sparsity property, meaning that only a few variables belonging to a few groups are effectively relevant, the Group Lasso provides a sparse solution, at a group level: within a group, either all of variables, or none of them, are selected. Oracle inequalities have been established, for the Group Lasso estimator, for prediction and estimation errors based on restricted eigenvalue or coherence assumptions [Bach, 2008, Lounici et al., 2011, Yuan and Lin, 2006]. The Restricted eigenvalue conditions imply, in the Group case, to compute minimal eigenvalue of sub normalized Gram matrices. As in standard HD linear model, checking those conditions is (presently) computationally intractable.

Under "strong group sparsity" assumptions, meaning that a small number of groups of "reasonable" size contains all the relevant variables and under group sparse eigenvalue conditions, Huang et al. [2010] show that the Group Lasso may be superior to the standard Lasso. Lounici et al. [2011] demonstrate that the Group Lasso can even provide smaller prediction and estimation errors than the Lasso.

3.2 Thresholdings with groups

As the LOL procedure is a counterpart of Lasso or Dantzig algorithms for ordinary sparsity, we introduce with D. Picard and K. Tribouley, the Group LOL procedure, which takes into account a group structure in the parameter estimation [A9], [A10]. As LOL, the Group LOL is a two-step **blockwise** thresholding procedure with no optimization. The group LOL procedure has been discussed in [Meinshausen, 2013], [Obozinski, 2013], [van de Geer, 2013], and [Yuan, 2013] and Mairal and Yu [2013] have provided an evaluation on the use of the procedure on a real application.

Structured Coherence

To handle groups and tasks, a (re) indexation of the columns of the design matrix X is provided. $X_{(j,t)}$ denotes the variable registered in the j^{th} group $\mathcal{G}_j$ for position (task) t. The design matrix is here supposed to be normalized such that $\frac{1}{n}\sum_{i=1}^{n} X_{i,\ell}^2 = 1$, $1 \leq \ell \leq k$. $\Gamma = X^t X$ denotes the Gram matrix of X.

Groups may have different sizes and $t_j = \#(\mathcal{G}_j)$ defines the cardinal of the group $\mathcal{G}_j$, $1 \leq j \leq p$ with $\sum_j t_j = k$.

To characterized the correlations of the "underlying" task and group structures, two coherence indices depending either on groups (denoted by γ_{BG}) or tasks (denoted by γ_{BT}) are introduced.

- The coherence between different tasks, with no restriction on the group membership is defined by:

$$\gamma_{BT} := \sup_{(j,j')\in\{1,\dots,p\}^2} \sup_{t\in\{1,\dots,t_j\},t'\in\{1,\dots,t_{j'}\},t\neq t'} \left|\Gamma_{(j,t)(j',t')}\right| \tag{3.4}$$

- The coherence for the same task but between different groups is defined by::

$$\gamma_{BG} := \sup_{(j,j')\in\{1,\dots,p\}^2,j\neq j'} \sup_{t\in\{1,\dots,t_j\wedge t_{j'}\}} \left|\Gamma_{(j,t)(j',t)}\right|. \tag{3.5}$$

Figure 3.1 illustrates the computation of γ_{BT} and γ_{BG}, on an example.

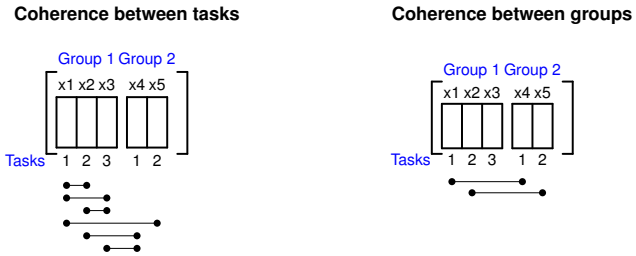

Figure 3.1: Illustration of the computation of coherence indices between tasks (γ_{BT} left) and groups (γ_{BG} right) for a design matrix with $k = 5$ variables spread in $p = 2$ groups.

The coherence of the design matrix X (denoted by γ in this Chapter) is then "split" in two indices such that $\gamma = \gamma_{BT} \vee \gamma_{BG}$. In the multi-task or in the no group case, we may observed that $\gamma_{BT} = 0$.

3.2.1 Learning Out of Leaders algorithm with groups

The Group-LOL procedure is presented, as for the LOL procedure, with the help of a pseudo code with the group structure, $\mathcal{G}$, added as input to the procedure.

$$(\hat{Y}, \hat{\beta}) \leftarrow \text{GroupLOL}(X, \mathcal{G}, Y, \lambda_1, \lambda_2)$$

Table 3.1: Definition of GroupLOL procedure: input= $(X, \mathcal{G}, Y, \lambda_1, \lambda_2)$, output= $(\hat{\beta}, \hat{Y})$.

The parameters λ_1, λ_2 define the levels of blockwise thresholds which play a similar role as the tuning thresholding parameters in the LOL procedure.

• **Initialization:**
An upper bound of the number of groups, N^*, that may be lately selected during the procedure, is first computed regarding the values of the index $\tau^* = t^*\gamma_{BT} + \gamma_{BG}$ and the precision parameter, ν, of the procedure where $t^* = max_{j=1\dots p} t_j$ corresponds to the maximal size of the number of variables in the groups.

$N^* = \nu/\tau^*$ *upper bound for the number of the group leaders,* $\tau^* = t^*\,\gamma_{BT} + \gamma_{BG}$

Table 3.2: GroupLOL initialization part.

The term τ^* is an essential quantity which will appear in the rate of convergence of the estimated coefficients (see Theorem 8). In the no group case, $\gamma_{BT} = 0$, the index τ^* is just the empirical coherence of the design matrix. In this case, the upper bound N^* is also the same as in the LOL procedure ($\gamma = \gamma_{BG}$, $\tau^* = \gamma$).

• **Step1: Find the group leaders by thresholding.**
The Group-LOL procedure selects at most $\mathcal{B}$ groups (and consequently the corresponding predictors inside each selected groups) which are "globally" the most correlated to the target regarding the "amount" of correlation brought by the overall predictors in each group and the value of the blockwise threshold λ_1.

$R_\ell = \sum_{i=1}^n Y_i\, X_{i\ell}$	predictor "correlation", $\ell = 1 \dots k$
$\rho_j^2 = \sum_{t=1,\dots,t_j} R^2_{(j,t)} := \|R\|^2_{\mathcal{G}_j,2}$.	Group $\mathcal{G}_j$ overall correlation, $j = 1, \dots, p$,
$\mathcal{B} = \left\{ j = 1, \dots, p,\ \rho_j^2 \geq \left(\rho^2_{(N^*)} \vee \lambda_1^2 \right) \right\}$	with $\rho^2_{(1)} \geq \dots \geq \rho^2_{(j)} \geq \dots \geq \rho^2_{(p)}$
$\mathcal{G}_\mathcal{B} = \cup_{j\in\mathcal{B}} \mathcal{G}_j$	Group leaders

Table 3.3: "Find the group leaders" by blockwise thresholding (GroupLOL/step1)

The set $\mathcal{G}_\mathcal{B}$ corresponds to the set of predictors selected at the end of step 1, for the $\mathcal{B}$ groups. If the first tuning parameter λ_1 is chosen such that $\lambda_1^2 > \rho_1^2$, $\mathcal{B}$ is empty and, in this particularly case, all the estimated coefficients are equal to zero $\widehat{\beta} = 0$.

• **Step2: OLS.** Group-LOL regresses the target on the set of predictors ($\mathcal{G}_\mathcal{B}$) belonging to the group leaders (Table 3.4).

• **Step3: Block thresholding.**
A second blockwise thresholding is applied on the estimated coefficients where λ_2 is the second thresholding parameter (Table 3.5)
The Group LOL procedure provides similar results to the Group Lasso. All the coefficients corresponding to not selected groups in step 1 or step 2 are tuned to zero. In the no group case ($k = n$, $t^* = 1$),

$\widehat{\beta}(\mathcal{B}) = [X^t_{\mathcal{G}_\mathcal{B}} X_{\mathcal{G}_\mathcal{B}}]^{-1} X_{\mathcal{G}_\mathcal{B}} Y.$	$\widehat{\beta}$ estimation
$\widehat{\beta}_{\mathcal{G}_\mathcal{B}} = \widehat{\beta}(\mathcal{B})$	
$\widehat{\beta}_{\mathcal{G}^c_\mathcal{B}} = 0$	$\mathcal{G} = \mathcal{G}_\mathcal{B} \oplus \mathcal{G}^c_\mathcal{B}$

Table 3.4: Ordinary Least Square on the Group leaders (GroupLOL/step2)

$$\widehat{\beta}^*_\ell = \widehat{\beta}_\ell \, \mathbb{I}\{ \|\widehat{\beta}\|_{\mathcal{G}_j,2} \geq \lambda_2 \} \quad \forall \ell = (j,t) \in \{1,\ldots,k\}$$

Table 3.5: Last blockwise thresholding on the estimated coefficients (GroupLOL/step3)

the Group LOL is similar to LOL procedure.

3.2.2 Main theoretical results

To guarantee the consistency of the Group LOL procedure, the following key assumptions are defined on the design matrix X (a1:a2), on the β coefficients (a3), on the structure of tasks and groups $\mathcal{G}$ (a4) and on the noise ε (a5):

(a1-a2) *Homogeneity and normalization of the design matrix* is assumed. For simplicity, we suppose here that we have the same number n of available observations across the predictors: $\sum_{i=1}^n (X_{i,\ell})^2/n = 1$, for $1 \leq \ell \leq k$.

(a3) *Group sparsity* condition on the "size" of the unknown coefficients is defined by a combination of ℓ_q between-blocks with ℓ_1 block norms. We then assume that there exist $q \leq 1$ and $M > 0$ such that:

$$\sum_{j=1}^p \|\beta_{\mathcal{G}_j}\|^q \leq M^q. \tag{Ag4}$$

(a4) The *precision* ν of the Group LOL procedure ($\nu \in]0,1[$) is supposed to be greater that the index τ^* equal to $t^* \, \gamma_{BT} + \gamma_{BG} \leq \nu$.

(a5) *Conditions on the noise:* ε is a vector of i.i.d. variables $\mathcal{N}(0,\sigma^2)$. A sub-Gaussian distribution may be also considered with zero mean and variance σ^2.

It should be underlined that most of these assumptions are easy to check on the data, especially compared to RIP conditions on sub normalized Gram matrices as for the group Lasso.

Regarding the previous assumptions (a1-a5), Theorem 8 provides the convergence rate of the estimation of the coefficients $\widehat{\beta}$ computed with the Group-LOL procedure when the threshold parameters λ_1, λ_2 are properly chosen [A9]. The values of the thresholds λ_1 and λ_2 depend on the design matrix (for the values of n and $\log(p)$), of the structure of tasks and groups (for the value of t^* and τ^*), on the β coefficients (for the M -group sparsity), and on the noise level(σ).

Theorem 8. *Assuming assumptions* $a1 - a5$, *we get:*

$$E\|\widehat{\beta}^* - \beta\|_2^2 \leq \square \left[\frac{t^* \vee \log p}{n} \vee (t^* \, \gamma_{BT} + \gamma_{BG})^2 \right]^{1-q/2}$$

where $\square$ *is a constant depending on* M, ν, σ^2, λ_1, λ_2.

Deduced from Theorem 8, the following key features have a direct impact of the sharpness of the procedure:

- *the architecture of the structure* induced by the term $\frac{t^* \vee \log(p)}{n}$
- *the correlation across tasks and groups* induced by the index $\tau^* = t^* \gamma_{BT} + \gamma_{BG}$.
- *the group sparsity of the vector* β characterized by $\sum_{j=1}^{p} \|\beta_{\mathcal{G}_j}\|^q \leq (M)^q$.

van de Geer [2013] shows that under coherence conditions, the Group Lasso achieves the same rate of convergence as the Group LOL, up to logarithmic terms.

3.2.3 Grouping versus no grouping

From an estimation point of view, it is easy to propose simple examples of HD linear models for which a specific group structure does not bring any benefits. For example, if we consider a model characterized by a small number of significant variables each of them scattered in different groups containing also many other variables with no significant coefficients, even if the group procedure may select some "significant" groups and variables, a huge number of no significant variables will be, at the same time, selected. In this specific case, it is obvious that the group structure will deteriorate the estimation. However, if an "appropriate" structure exists (which may be characterized by some group sparsity assumptions), [Huang et al., 2010] show that, better results compared to standard HD linear models can be obtained using a group algorithm, as for example the Group Lasso.

The following example explains and quantifies, for a specific case, the gain that can be expected from the Group LOL compared to the original LOL procedure [A10]. We consider a situation where the parameter β has ST non-zero coordinates which are all equal to γ and we consider $p = k/T$ groups of covariates with the same size T characterized by $\gamma_{BT} \sim 0$ and $\gamma_{BG} = \gamma\sqrt{\log(k)/n}$. The following table shows the rate of convergence when comparing grouping versus without-grouping, depending on the places of the contributing β coefficients in the different groups [A10].

	LOL	Group LOL/optimal case	Group LOL/worse case
Rates	$ST(\gamma^2 + \frac{\log k}{n})$	$S(\gamma^2 + \frac{T}{n} + \frac{\log k/T}{n})$	$ST(\gamma^2 + \frac{T}{n} + \frac{\log k/T}{n})$

- Group LOL/optimal case corresponds to the case where the non-zero coefficients are all gathered in S groups (of size T), while
- Group LOL/worse case corresponds to the case where they are scattered in ST groups.

This simple example emphasis that we gain in grouping by taking groups of relatively small size (less than $\log(p)/n$) . With no appropriate groups, the price to pay may be heavy compared to the no group case, especially when γ_{BT} is not zero and when the maximal size of groups, t^* is high.

Base on the previous results (Theorem 8) and on the possibility to improve the rate by a smart distribution of the predictors into the groups, we propose a data driven strategy to built the groups in order to boost the rate.

3.3 Data driven strategies to built relevant groups

In some applications, the structure of the groups is driven by some precise requirements. In this specific cases, compared to standard HD linear models, the objective of the Group Lasso or the Group LOL procedures is to estimate the coefficients of the variables belonging to the most relevant groups. In various cases, however, there is no obvious grouping at hand. In this direction of research, Zhao et al. [2009] proposed, to use a robust version of the K-means to group the features. Bühlmann et al. [2013] proposed a bottom-up agglomerative clustering algorithm based on canonical correlations.

To boost the rate of convergence of the Group LOL procedure (Theorem 8), we propose a data-driven strategy first to compute the appropriate number of groups then to fill the groups and tasks with the appropriate variables.

In order to introduce our boosting grouping strategy (BG), two preliminary ways of building the groups (called strategies) are first introduced relying on *gathering* either *scattering* the variables [A9] in the groups. An extensive simulation study illustrates, at the end of this section, the practical benefits of the BG strategy.

Gathering

Assuming that the number of groups (p) is known, a natural idea in order to design a predictive model with groups is to 'gather' the p covariables of the design matrix X regarding their "correlation" value with the target (for weak coherence assumptions). The "Gathered Grouping" **(GG)** strategy gathers the variables exhibiting similar absolute correlation values with the target Y. The p different groups are then filled by using the ordered indices:

$$\mathcal{G}_1 = \{(1), \ldots, (\lfloor k/p \rfloor)\}, \quad \ldots \quad , \mathcal{G}_p = \{(k - \lfloor k/p \rfloor), \ldots, (k)\}$$

where (ℓ) denotes the index associated to the ranking quantity $|R_{(\ell)}|$ where $R_\ell =< X_\ell, Y >$, $1 \leq \ell \leq k$.

In low coherence cases, the index value brings a prior information about the significativeness of the co-variables. This approach let to built groups characterized by a strong group sparsity [Huang et al., 2010] (a few groups are effectively relevant).

At the opposite, without taking any care on the correlation between the target and the predictors, the Gathering at Random (GR) strategy gathers in each group k/p randomly chosen variables among the k initial regressors (without replacement). On the opposite to the GR strategy, this last strategy does obviously not bring any clever groups.

Scattering and boosting the rates with the Group LOL

In order to increase the rate of convergence, the following quantity which only depends on the arrangement of the co- variables into the groups (and not on the target) has to be smaller as possible:

$$\sqrt{\frac{t^* \vee \log p}{n}} \vee \{t^* \gamma_{BT} + \gamma_{BG}\} \tag{3.6}$$

We propose a "Boosting rate with Grouping" strategy (called BG), which addresses, at the same time, the question of choosing the number of groups (p), as an "optimal" repartition of the predictors inside the groups. The BG strategy first selects, in the initial design matrix, the most p^* correlated predictors (step1), then scatters them into p^* groups taking care to decrease as much as possible the γ_{BT} index (without increasing $\log p/n$) (step2). The remaining predictors are finally gathered in the different tasks and groups (step3), taking care of the group sparsity assumption. The three steps are detailed hereafter:

BG (step 1/3): Determination of the number of groups

The overall correlation of predictors are sorted in descending order such that:
$\gamma = \max_{\ell>\ell'} |\Gamma_{\ell\ell'}| \geq \ldots \geq |\min_{\ell>\ell'} \Gamma_{\ell\ell'}| \geq 0$.

The cardinal, denoted by $p(u)$ of the set of predictors $\mathcal{D}_u$, characterized by a correlation higher than γ/u with $u > 0$ is computed:

$p(u) = \#\mathcal{D}_u$ with $\mathcal{D}_u = \{\ell \in \{1,\ldots,k\}, \exists \ell' \in \{1,\ldots,k\} \setminus \{\ell\}$ such that $|\Gamma_{\ell\ell'}| > \gamma/u\}$.

In the BG strategy, the appropriate number of groups p^* satisfies $p^* = p(u^*) = \lfloor k/u^* \rfloor$, where u^* gives the size of groups (for simplicity, we consider that k/u^* is an integer).

Figure 3.2 illustrates the evolution of the function $g(u) = k/u$ and $p(u) = \#\mathcal{D}_u$ for i.i.d. Gaussian variables, $n = 200$, $k = 1000$.

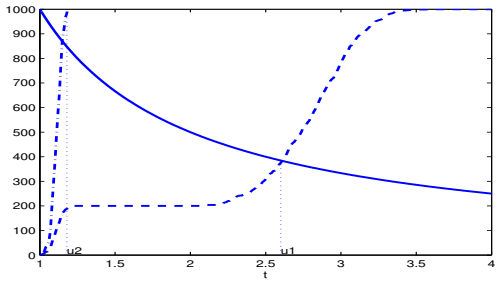

Figure 3.2: X-axis: Common size t_1. Y−axis: number p of groups. Solid line: $g(u) = k/u$. Dashed line: $p(u)$ for $\rho = 0.5, \pi = 20\%$ (see simulation part). Dot dashed line: $p(u) * \log p(u))$. Dot lines: corresponding u_1, u_2 positions. $n = 200$, $k = 1000$, SNR $= 5$.

BG (step 2/3): Scattering the most correlated predictors across the groups

The p^* elements of $\mathcal{D}_{u^*}$ are affected to the task "number one" of each group. This repartition provides $\gamma_{BT} = \gamma/u^*$, $t^* = u^*$, $\gamma_{BG} = \gamma$ and consequently $t^*\gamma_{BT} + \gamma_{BG} \leq 2 * \gamma$

As soon as $\gamma \geq c[\log p/n]^{1/2}$, under the appropriate assumptions $a1 - a5$, Theorem 8 provides for the BG strategy:

$$E\|\hat{\beta}^* - \beta\|_2^2 \leq \square\ (\gamma)^{2-q}\,. \tag{3.7}$$

BG (step 3/3): Gathering the predictors

Before the completion of the groups, the groups are re arranged by sorting the correlation indicators associated to the delegates: $R_{(1)} \geq \ldots \geq R_{(p^*)}$. This means that $\mathcal{G}_1$ contains the delegate ℓ_1 such that $R_{\ell_1} = Y^t X_{.\ell_1}$ takes the largest correlation value (equal to $R_{(1)}$) and $\mathcal{G}_{p^*}$ has the delegate with the smallest $R_{\ell_{p^*}}$ correlation value (equal to $R_{(p^*)}$). The groups are then built such that the R's are as homogeneous as possible in each group and as close as possible to their delegate. Grouping starts by ranking the remaining R's (i.e. not associated to a delegate): $R_{(1)} \geq \ldots \geq R_{(k-p^*)}$. The p^* different groups are then successively filled by using the ranking indices:

$$\mathcal{G}_1 = \{\ell_1, (1), \ldots, (\lfloor k/p^* \rfloor) - 1\}, \quad \ldots \quad , \mathcal{G}_{p^*} = \{\ell_{p^*}, (k - p^* - \lfloor k/p^* \rfloor) + 1), \ldots, (k - p^*)\}.$$

Without taking any care on the correlation between the target and the predictors, The Boosting at Random strategy (BR), fill the groups completed randomly (except for step 2): the $k - p^*$ variables are spread out randomly into the p^* groups.

3.4 Numerical experiments and applications

Extensive simulations were conducted to explore the benefits of the different grouping structures using the Group LOL procedure [A9]. We just recall the experimental design as the main results.

3.4.1 Experimental design

The design matrix X is a standard Gaussian $n \times k$ matrix. Each column vector $X_{\cdot\ell}$ is centered and normalized, $1 \leq \ell \leq k$. The target observations Y are computed using $Y = X\beta + W$ where β is a vector of size k whose coordinates are zero except for S coefficients which are equal to $\beta_\ell = (-1)^{b_\ell}|z_\ell|$ for $\ell = 1, \ldots, S$ where the b's are i.i.d. Rademacher variables and the z's are i.i.d. $\mathcal{N}(5, 1)$ variables. ε are i.i.d. variables $\mathcal{N}(0, \sigma^2)$. The variance σ^2 of the noise is chosen such that the SNR (signal over noise ratio) is close to 5 which corresponds to a middle noise level. To introduce some dependency between the regressors, we chose randomly $\lfloor \pi k \rfloor$ variables among the k initial regressors ($\pi = 5\%, 10\%, 20\%$), which are the characterized by a mutual correlation equaled to ρ ($\rho = 0.0, 0.6, 0.8$). This method has the advantage to tune accurately the number of correlated variables as well as the amount of correlation between the variables.

Benefits of boosted grouping

The performances of the different grouping strategies are presented in Figure 3.3 for different sparsity levels S and different levels of dependence (ρ, π). For each strategy (either GG, GR, BG, BR strategies as defined previously), the relative prediction error $E_Y = \|Y - \hat{Y}\|_2^2 / \|Y\|_2^2$ is computed on the target Y.

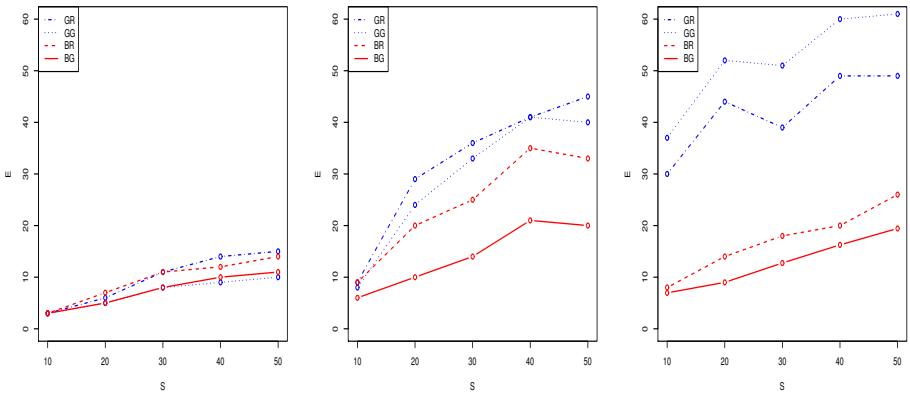

Figure 3.3: Performances (relative prediction error) computed for different grouping strategies (average over $K = 100$ repetitions). Left: no correlation between predictors, $\pi = 0$, $\rho = 0$. Center:$\pi = 0.2$, $\rho = 0.6$. Right: strong rate of correlation and high correlation values, $\pi = 0.4$, $\rho = 0.8$

We observe that the Boosting Grouping strategy always shows the lowest relative prediction error, for the different cases. The BG strategy takes especially advantages when strong correlation are observed in the design. For instance, when ρ and π are significatively high ($\rho = 0.8$ and $\pi = 0.4$), the boosting procedure clearly shows substantial benefits as illustrated in Figure 3.3. In the no-dependency case ($\pi = 0$), when the sparsity is high ($S = 10, 20, 30$), similar performances are obtained for any grouping strategy which seems clearly understandable.

Comparison with the Group Lasso

The Group-LOL procedure associated to the Boosting Grouping strategy (BG) has been compared with the Group Lasso. Both group procedures are built using the boosting strategy (BG) and cross-validation are both used to determine the final model.

Comparison of the prediction results show similar behaviors when there is no high correlation between the co variables ($\pi = 0$) or when the sparsity ($S = 50$) is small. In the other cases (especially when the sparsity is large i.e. S small), Group-LOL always outperforms the group lasso [A9]. To end this comparison, we should add a few words about computational aspects. Regarding the complexity of the different methods, Group-LOL has a strong advantage over the Group Lasso. The Group Lasso algorithm is based on an optimization procedure which can be time consuming while Group-LOL procedure solves the penalized regression using two thresholding steps and a classical regression.

3.4.2 Catching feature with the grouping strategy

Mairal and Yu [2013] have investigated the Group LOL procedure for neuroscience data and proved that the method shows satisfactory performances in this context but was, in their case, not especially suitable in catching interesting features. The following example shows that this 'can' nevertheless happen [A10].

Going back to fit the global electrical consumption in France using high dimensional sparse methods as introduced in Chapter [2] the variable Y of interest is a daily electrical signal recorded each half hour

presented in Figure 3.4. Such intraday load curves can be explained using both types of variables: climate variables are essential (basically temperature recorded on the same day at different spots in France) and these curves show also typical shape features which are generally well captured using dictionaries of functions such as wavelet bases, the Fourier basis or combination of both types of dictionaries, with the serious issue that these dictionary functions happen to be highly correlated with the climate variables and very often disappear in sparse representations.

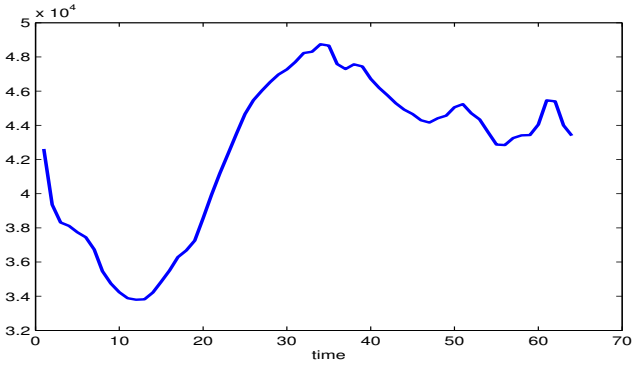

Figure 3.4: Electrical consumption signal

We consider the following dictionary $\mathcal{D} = \{E, C, S, H, T\}$ composed of a family of $k = 114$ heterogeneous functions including a set of climate functions recorded all over France during the same day as illustrated in Figure 3.5 and a set of generic shape functions from the trigonometric and Haar bases:

- E is the constant function: $E(t) = I\{[0,1[\}(t)$,
- $C = \{C_1, \ldots, C_{31}\}$ are the cosine functions with increasing frequencies:

$$C_\ell(t) = \sqrt{2}\cos(2\pi(2\ell - 2)t)$$

- $S = \{S_1, \ldots, S_{31}\}$ are the sine functions with increasing frequencies:

$$S_\ell(t) = \sqrt{2}\sin(2\pi(2\ell - 1)t)$$

- $H = \{H_2, \ldots, H_{31}\}$ are the Haar functions with increasing frequencies:

$$H_\ell(t) = \psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k) \quad \text{where} \quad \psi = I\{[0,1/2[\} - I\{[1/2,1[\}$$

- $T = \{T_1, \ldots, T_{20}\}$ are the 20 temperature functions recorded at different spots represented Figure 3.5.

The BG procedure is used to organize the functional predictors into different groups. $p = 35$ delegates emerge from the dictionary $\mathcal{D}$. The set of delegates is composed of all the 20 temperature (T), the constant (E), 3 cosine functions, 4 sine functions, and 7 Haar wavelets. All the 15 generic functions (constant, cosine, sine, and Haar) selected in the delegate set are strongly correlated to the temperature

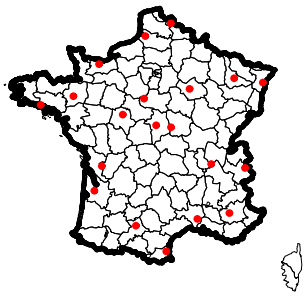

Figure 3.5: French temperature spots

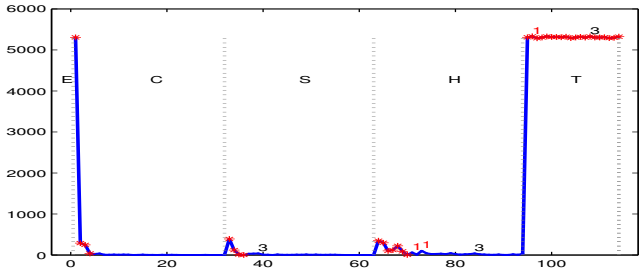

Figure 3.6: "Correlation" between the consumption signal and the various dictionary functions. The chosen delegates, for LOL procedure, are tagged with a red star.

signals and are described in Figure 3.6. At this step, we observe that the set of delegates brings meaningful patterns and mostly catch the climate information, which is known to have a high impact on the electrical consumption.

The 79 ($k - 35$, $k = 114$) remaining functions are then gathered in the $p = 35$ groups following the repartition rule of BG procedure. Each group is then defined by 3 or 4 functions. For illustration, Figure 3.6 gives the composition of two groups. The first group is composed of one Temperature and 2 Haar functions (tag '1', Figure 3.6), second group of one Temperature and one sine function and one Haar function (tag 3). As expected, the coherence for the tasks is weak ($\gamma_{BT} = 0.35$ and $\gamma_{BG} = 0.99$). Note also that this step in not only depending on the dictionary but also incorporate information on the signal Y. In order to compare the benefits of grouping versus non grouping, the LOL procedure is also performed on these data.

When Group LOL is used, the relative prediction error equals 0.75%. 24 regressors allocated among 8 groups are requested: `THS-THH-THH-TCS-TCST-HHTC-STSH`. Hence we find at the end as meaningful functions 8 temperature ('T'), 3 cosine ('C'), 5 sine ('S') and 8 haar ('H') functions.

For LOL, we impose the same number (24) of selected functions as for Group LOL to induce fair comparison. In this case, the selected functions computed by LOL are:

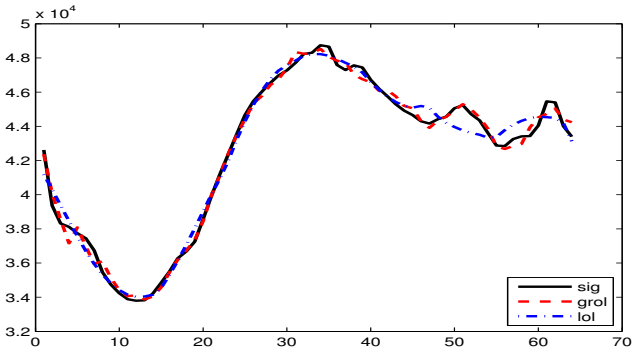

Figure 3.7: Model of the consumption signal (black-solid line) using Group LOL (red-dahsed line) and LOL (blue dot dashed line).

`T-T-C-T-E-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-S-C` described by 20 temperature ('T'), 2 cosine ('C'), 1 sine ('S') and the constant ('E') functions. LOL selects, as expected, all the functions strongly correlated with the target consumption signal. Nevertheless, these functions are also strongly correlated which each others and bring redundant information. The LOL relative prediction error equals 1.86%, which is 2.5 time greater than Group LOL relative error.

Figure 3.7 shows the consumption signal and the fitted signals computed either with Group LOL or with LOL procedure. Notice that one big benefit of the BG procedure is to impose diversity and as a consequence, Group LOL algorithm selects different families of functions (temperature, cosine, sine, haar, constant) which improves seriously the fitting result. We agree with Mairal and Yu [2013] that the groups formed on real data by an automatic procedure, as the BG procedure, are not always easily interpreted by the experts of the field. Nevertheless, as it is the case here, imposing diversity as a key principle may be helpful for the interpretation and may also induce a better precision.

3.5 Software

A software package was developed for the Group LOL procedure and published [S6]. This package was used by J. Mairal & B. Yu on an experiment neuroscience data set. They conclude that the "Group-LOL method performed relatively well given complexity and data [Mairal and Yu, 2013].

3.6 Conclusions

This experimental study shows that true benefits can be obtained using a grouping approach for HD linear models even in the case where there is no prior knowledge on the groups. However, the results are highly relying on the grouping strategy. The boosting strategy brings a satisfactory answer to the grouping problem when no prior information is available on the structured sparsity. This strategy is very easy to implement and especially well adapted when a strong correlation exists between the regressors in the case of high sparsity.

Chapter 4

Industrialization of statistical or machine learning algorithms

I started the work presented in this chapter, in october 1999, when I left my position at Paris X University to participate in the creation of Miriad Technologies, a private company specialized in the elaboration of decision making softwares based on operational data. I had then the opportunity to be involved in many industrial projects and to work in a scientific collaboration with R. Azencott, J. Besnard, B. Durand, O. Cherif, F. Gautier, K. Fakhr-Eddine, O. Gérard, J. Lacaille, J.F. Legrand, A. Maazi from Miriad Technologies (and many others...). Most of these projects, as I already mentioned in my career overview, were developed under non disclosure agreement and were not published in academic journals. However, technical reports were systematically written and delivered to the clients [Anx: Miriad Tec. Reports].

When I took my academic position back, similar works were provided with R. Azencott and J. Wang at Centre de Mathématiques et de Leurs Applications at Ecole Normale Supérieure de Cachan, in the European projects **ADHER** (Automated Diagnosis for Helicopter Engines and Rotating parts, Eu 030907), **Innotex** (INNOvation within the TEXtile manufacturing lines in Europe, Eu 030312) and **TRACE** (TRAffic Causation Analysis in Europe). Parts of this work have been published in:

C5 O. Cadet, C. Harper, M. Mougeot (2005) *Monitoring Energy Performance of Compressors with an innovative auto-adaptive approach.* Instrumentation System and Automation -ISA- Chicago.

E2 R. Azencott, J.P. Kreiss, M. Mougeot, P. Pastor, M. Pfeiffer, S. Siebert, T. Zangmeister. (2007) *Analysis Methods for Accident Causation Studies.* TRACE European project No. 027763.

E11 ADHER (2009) *Automated Diagnosis for Helicopter Engines and Rotating parts* Publishable final activity report (2009), ADHER FP6/Aeronautics project AST5-CT-2006-030907.

E1-E10 see European Project section

—

4.1 From research to development: step by step

When I was working at Miriad Technologies or when I was working as an associate professor, I have initiated many collaborations with industrial partners. Most of these collaborations were oriented towards the research and development of innovative solutions, relying on statistics or machine learning

methodologies. The final aim was always to be *u*sed in an operational environment through a software. To succeed in this objective (deployment in an operational environment) and according to my former experiences, I noticed that this kind of projects always followed the same successive tasks:

1. interviews before project (task 0),
2. Proof of concept (POC) (task 1),
3. pilot software (task 2),
4. industrial software (task 3).

It should be underlined that Go/NoGo decisions always end each task. To be honest, for various reasons, most of the projects end after the first POC task (due to financial, technical, human reasons).

Having the opportunity to perform all successive tasks may be seen as a success.

4.1.1 Before the project

Before being able to effectively start a collaboration, a necessary task for the researcher, who will be potentially involved in the future project, is to deeply understand the operational requirement, to translate the operational need into a statistical or mathematical question, to sketch a first solution, to be able to give some clues, and finally to convince the industrial partner that the solution will meet his requirement. This preliminary work needs to be performed before any data analysis or before any contractual collaboration, in a relatively short period of time, and this is often a challenging task! In order to achieve this first phase, short technical interviews are conducted to ensure that a solution could be provided. Sometimes, alternative solutions need to be imagined and proposed as it does not seem possible to answer directly to the first operational need.

According to my former experience, this tasks needs to be take in charge by a Researcher (or a former Researcher) and a Seller. The Miriad Technologies experience has shown that without deep scientific skills, it was impossible to sell any Proof of Concept.

4.1.2 Proof of concept (POC)

A collaboration always starts with, what is called, a *proof of concept* (POC): regarding the industrial requirement and a set of historical data, a *methodology* based on mathematical or statistical tools, is elaborated to prove the concept. As mentioned by [Breiman, 2001] and [Wickham, 2014], this work needs to:

- "Live with the data before you plunge into modeling",
- Search for a model that gives a good solution, either algorithmic or theoretical,
- evaluate the prediction accuracy on test data sets to characterize how good the model is.

For all these tasks, the computer is an indispensable partner associated with more or less elaborated programming languages such as Matlab, Python,R, SAS, SPPS... with statistical or machine learning packages.

Of course, the development of a POC needs statistical knowledge. However, all tasks corresponding to a better understanding of the data are also essential. For example, interviews with human experts to better understand the process, exploration of the data base to get tidy data, visualization of the data to get intuitions on the modeling are also very important.

The development of a POC is never a straight line from raw data to a model. Exploratory data analysis and modeling tasks are, most of the time, very imbricated. At the end of the POC, communicating the

results is an essential step. If the industrial partner does not understand nor "feel" the solution, he will not believe in its use in an operational environment, and the story will end just after the POC!

Exploratory analysis and modeling clues.

An initial work, for which people do not have often many considerations but which is a necessary part of the project is to *"check" the statistical value of the data set*. During this task, very basic treatments may be extremely useful to exhibit potential data inconsistency, before any modeling. When the data set corresponds to a deep historical period, it is common to observe storage modifications such as unit changes or sensor modifications.... Some basic rules are usually defined, during the first preprocessing of data cleaning, to keep or to discard numerical values or variables. When the number of variables is not too large, up to 100 (for instance), the visualization of the empirical distributions of the variables always brings an added-value, to check the data, but also to get some clues and some intuition on the potential underlying statistical model. Segmentation or quantification may follow this step to extract relevant behaviors regarding multi modal distributions.

Modeling.

The final mathematical or statistical solution is never given before a project. Both problem and data progressively lead to the solution. During a POC, starting from data, we have to imagine and provide an answer to the operational needs. The best available solution to a data problem might be either a stochastic model or an algorithmic model. Programs are developed to apply the methodology to a set of operational data and to evaluate the performances of different models. Very often, a bunch of models (parametric/non parametric) are tested, and the criterion to compare the model is often the predictive accuracy and complexity plays a major role.

Software tools.

At Miriad Technologies, during 6 years, we developed a proprietary Rapid Application Development (RAD) called Miriad Process which has been progressively enhanced with the different methodologies introduced, developed and programmed during the successive Proof of concepts [Lacaille, 2003]. The Miriad process tool helped us to perform proof of concepts more rapidly. The former tool used in Miriad can be today compared to SAS entreprise miner®.

Today, I use Matab, R or Python for the development of the POC, depending on the data, and depending on contractual specifications. The R language, which does not provide ascendant compatibility, is, according to my experience, not always welcomed in some companies.

Deliverable.

The deliverable of the POC is always a report describing the method and the associated performances computed, with different models, on test data sets.

POC illustration.

At Miriad Technologies, most of the industrial customers were interested in the design and the evaluation of automatic monitoring systems [MiriadTecReports]. Among many others, POC were conducted for electricity monitoring [Mougeot, 2000], welding anomaly detection[Mougeot and Fakhr-eddine, 2002], chemical reactors [Mougeot and Layeillon, 2004, Mougeot and Maazi, 2000, Mougeot et al., 2003], and equipment monitoring [Mougeot, 2003, 2005],

4.1.3 Pilot

Given the results of a POC, and if the industrial partner can expect some benefits from the industrialization of the method, the development of a pilot software is usually the next step. The statistical methodology is then implemented into a prototype software. The algorithmic solution, developed during the POC period, is upgraded to be used in an industrial environment and to offer a more robust behavior.

The statistical methodology is packed into a "component" (compiled or not), and data base connections (such as Human Machine Interfaces) are added and developed. During this phase, a site of industrial production is carefully chosen *to follow, in an operational environment and over a given period of time*, the performances of the methodology. This well help to evaluate precisely the benefits and the return on investment of the method and sometimes to compare it to existing "home made" solutions.

4.1.4 Industrial software

When the use of the prototype shows added value in the operational context, a software package may be developed, then integrated into the IT system of the industrial partner to be finally deployed on other industrial sites of production for decision making processes. Whereas a POC requires exclusively statistical skills, the two last steps (pilot and deployment of an industrial software) require more computational abilities. *It should be stressed that feedbacks from the operational field always bring relevant information regarding the behavior of the method in a real environment and systematically raise new methodological questions and technical points which must be analysed to improve the solution.* It is often necessary, during the two last phases, to update the mathematical or statistical methodology developed during the first research phase to assess, in an operational context, the first performances obtained off line. From my point of view, to ensure a win to win global project, statistical skills must be involved until the end of an innovative project using machine learning or statistical knowledge and not replaced directly after the POC by exclusively computational abilities.

It should be stressed that collaborating with an industrial partner implies to be able to communicate throughout the project period, and to able to bring technical elements during all steps, in order to show that the current R&D has an added value for the company. According to my former experiences, a large majority of projects ends after the first POC step: most of the time because the evaluated performances do not convince the partner of any added value or because they do not reach the profitability target.

The following sections detail two success stories of valorization of research initiatives, I am particulary proud of. Both applications concerned sensor based health monitoring for industrial equipment with predictive modeling and were transformed into software packages.

4.2 Embedding R&D in Software

The first application deals with the "diagnosis of over consumption for compressors" (*SPEC+*), and the second one with "Automated Diagnosis for Helicopter Engine and Rotating parts" (*ADHER*).

4.2.1 Monitoring Energy Performance of compressors

A decision support software package for detecting excessive power consumption on individual compressors has been developed as a Miriad Technologies R&D project from 2001 to 2004 [C5]. The project successfully followed with success the three successive phases: POC (1)/ Pilot (2)/ Deployment (3). Large compressors are high energy consumers and critical elements for the production and distribution of air gases (oxygen, nitrogen,argon...). The monitoring of the health of these equipments is essential to be able to guarantee the production of gas and to control the energy costs. Online monitoring helps to optimize maintenance operations and to avoid sudden break-downs. The development of the SPEC+ project was characterized by three steps:

1. A Proof of concept was first dedicated, off line, to the design of a statistical methodology to automatically diagnose the over consumption of compressors and to evaluate the prediction accuracy

on a set of historical data.

2. A pilot: a software component was developed then installed to monitor 5 compressors, during 6 months, on an operational site (Air Liquide/France/Dunkerque) and to evaluate the potential return of investments.
3. The SPEC+ software package, including Human Interface Machine (HMI), Data base connection and the R&D component, was finally developed and installed in the Operational Control Center (Air Liquide America/Houston/US) to supervise more than 50 compressors, spread among 10 production plants on the Gulf coast [C5].

The statistical method introduced to monitor the compressors are presented hereafter. The method is generic and can be used to monitor other types of equipments. Points, specifically linked to compressor equipment are voluntarily discarded.

From predictive modeling to diagnosis

A full instrumented compressor owns sensors which record continuously up to 6 measures: electrical consumption, flow, input pressure, output pressure, water temperature and gas temperature. All these measures are not always available and depend on the level of instrumentation of each compressor in the plants. The Supervisory Control and Data Acquisition system (SCADA) let to retrieve periodically data for each compressor (for example, hourly sample rate). The aim of the SPEC+ POC was to design a method to be able to diagnose power over consumption on individual compressors. For this project, the method was defined in two steps: first the monitoring of the electrical consumption using predictive models then the diagnosis of potential over consumption based on a stochastic regression model. For each variable, a light preprocessing defined by a bandpass filter was, as usual, introduced to handle measurement errors.

Equipment modeling. For each compressor, we propose to introduce a stochastic regression model to explain the electrical consumption (target variable Y) regarding the other available contextual variables (X) by: $Y = g(X) + \varepsilon$. The function g is defined on the space of the available contextual variables and g(X) defines the part of the target variable which can be predicted by the contextual variables. ε is the residue of the decomposition and is considered as a random disturbance. Given an historical data set, D of (Y, X), automatic learning of the regression model comprises two successive parts: deterministic learning for the conditional density estimation ($\hat{g}_{\mathcal{D}}$) and random part estimation for the probability distribution of ε. As the prediction needs to be accurate, different methods of estimation of g need in practice to be investigated and evaluated on the data. Most of the time of the POC is usually devoted to the elaboration of an "appropriate" estimation of function g. In the SPEC+ regression model, g is a globally non linear function composed of different sub-linear models which are automatically defined and tuned given the empirical distribution of co variables. The parameters of the model should be estimated on a set of data, which are supposed to contain "no" over consumption.

Equipment monitoring and diagnosis. For a new observation (y_{new}, x_{new}), the monitoring of the equipment (here the compressor) is provided by the computation of the deviation between the observed and predicted target value $\hat{\varepsilon}_{new} = y_{new} - \hat{g}_{\mathcal{D}}(x_{new})$ and by the probabilistic evaluation of this deviation regarding the distribution of ε. A statistical fitting test is naturally proposed to diagnose the health of the equipment: testing the null hypothesis "usual healthy working equipment " against the alternative "non usual work" (which can be interpreted as no healthy behavior given the choice of the target variable). The diagnosis of the equipment is provided by the computation of the p value, $p_{new} = \mathrm{Proba}(\varepsilon > \hat{\varepsilon}_{new})$ and its comparison to a given level of risk, regarding the distribution of ε. Figure 4.1 presents the diagram of the methodology presented in [C5].

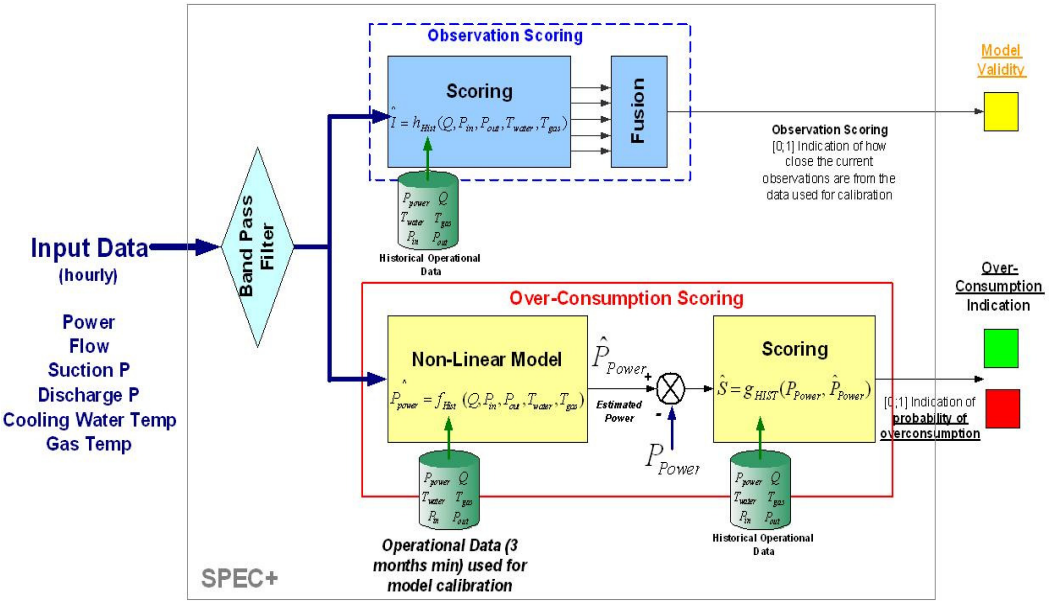

Figure 4.1: SPEC+ design as presented in [C5]

During the 6 months of this POC, a benchmark comparing the SPEC+ model and the existing home made models developed by Air Liquide was conducted by Air Liquide. The SPEC+ model, evaluated on 7 different compressors, showed the best performances with an average prediction accuracy below 1% which met Air Liquide requirements.

Monitoring the model

The monitoring method of consumption aims to be used through a supervision software in an Operational Control Center following many compressors (up to 50). For each equipment (compressor), a model is calibrated using a given set of historical data, and by consequence, each compressor diagnosis strongly depends on the underlying statistical distribution estimated, given the reference set of data. If the reference statistical distribution does not reflect the current work of the compressor, the diagnosis may be inappropriate. For example, it is well known that compressor efficiency depends on outside temperature: compressors are globally more efficient in winter than in summer. If the set of co variables does not contain any temperature information, a model calibrated with summer data will not be able to trigger any over consumption alarms in winter, because of a negative bias. At the opposite, such a model calibrated with winter data will continuously trigger alarms in summer, because of a positive bias. An important feature is then to be able to recalibrate the model with new data, if necessary. When implementing a self calibrated component in an operational field, it is essential to introduce in parallel an indication of the adequation between the current observations and the model (the reference set of data). For that purpose, a goodness of fit test can be introduced for the co variables comparing the current distribution of data used for diagnosis and the historical distribution. This is well known for a statistician, but, without such feature, the solution won't be able to be used over a long period of time, in an operational context. In the SPEC+ application, an input validity score was introduced to warn the operators of an inappropriate running model [C5].

The SPEC+ Software

Figure 4.2 shows two screenshots of the Human Machine Interface of SPEC+ software. In the left part of the figure, no diagnosis of over consumption was provided, however the model was diagnosed as inappropriate (yellow flag under the green light). In the right part, an over consumption alarm is triggered with an appropriate model.

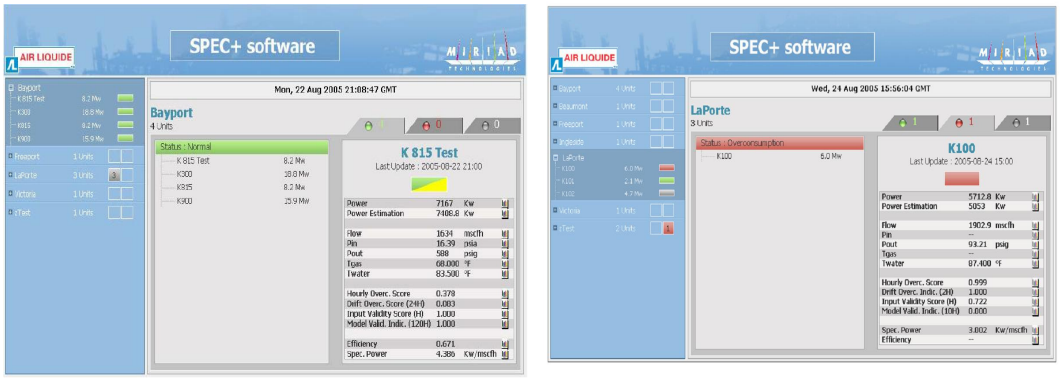

Figure 4.2: SPEC+ HMI.

Operational feed-backs

A first pilot program was launched in France in 2003, where the SPEC+ methodology was evaluated on line on an Air Separation Unit (ASU) for 6 months. The development of a full software package helped, afterwards in August 2005, to evaluate the benefits of the methodology in the USA on the Air Liquide Gulf Coast Pipeline plants. As part of Customer Acceptance Test, a series of tests were conducted on compressors in order to ensure that the software tool correctly detects online cases of over consumption. The performances of the software were directly evaluated by an Air Liquide team. In all cases, the software tool correctly detected over consumption on the compressor as there was no obvious indication available in SCADA that the compressor was in over consumption. The performances of the software fully met Air Liquide expectation. The good results from the validation field as well as an insight on implementation and validation of the final product were communicated by Air Liquid America at The Instrumentation, Systems and Automation Society [C5].

4.2.2 Automated Diagnosis for Helicopter Engines and Rotating parts

A second project of health monitoring I was involved in, was developed in the framework of the European project ADHER project from 2008 to 2009 and focus on "Automated Diagnosis for Helicopter Engines and Rotating parts". As for the SPEC+ project, the "heart" of the monitoring system relies on predictive modeling. Due to the nature of the data (vibration data), a huge amount of preprocessing were realized in order to implement sparse and efficient models. The ADHER software, developed in collaboration with J. Wang, successfully passed the two successive phases which are necessary for the valorization of a new methodology in the industry: from a Research to a Development project [E11]. The last phase (deployment) was not in the initial scope of the European project. ADHER project has been developed

under a non-disclosure agreement to ensure data and process confidentiality, but a global final activity report was published from the European consortium [E11].

To give a short introduction on the industrial background, helicopters availability, in-flight reliability, and low cost maintenance are major concerns for helicopter operators. Nowadays, accelerometers and shaft speed sensors are installed on helicopter critical components. Most of the data, recorded by on board sensors on engine and rotating parts, are systematically analyzed after each flight to provide monitoring of the equipment and to help diagnose potential failures as soon as possible. Monitoring systems are, most of the time, exclusively based on the analysis of the vibrations linked to shaft speed signals. In aeronautics, several contextual variables such as load, thermodynamics parameters or flight conditions are known to influence vibration regimes, and these contextual variables may take very different values between flights or even during flights. Methods to analyze vibration data taking into account contextual information are missing from most of these monitoring software based tools. Some of the main scientific and technological objectives of the ADHER project were to define innovative auto-adaptive algorithms enabling data-driven automatic learning to analyze empirical time evolutions of sensor data and to generate anticipative health diagnosis, taking into account context variables. These algorithms were tested on helicopter fleet vibration data to evaluate the feasibility of automated health monitoring of helicopter fleets.

Massive operational fleet data

The project database included data from 4 helicopters, recorded from 2004 to 2006. A total amount of 2000 flight hours were recorded through a set of discontinuous 10 sec intervals. This data base contains 45 000 flight records of 10 sec for the 4 aircrafts. In this data base, only 115 records corresponded to existing failures (0.2%). Each interval was characterized by vibration raw data recorded from 18 vibration sensors sampled at 48 KHz and by 6 contextual variables sampled at only 10 Hz. A strong difference of sampling rate was observed between vibration signals (48kHz) and contextual variables (10Hz). For a 10 sec interval, as 480 000 values were available for vibration, only 100 values were available for one contextual variable. A size of 100GB was necessary to store the whole initial data base size, which can be qualified as a relatively "Big data" data base.

Main features extraction

Before being able to implement any predictive model, a very large amount of work was dedicated to extract relevant features from vibration signals, according to physical mechanical properties Klein [2006]. For example, the vibration signals were first sampled according to the shaft speed, and power spectrum of the re-sampled signals were then computed. Estimation of energies at given spectrum pointers (♯20) were extracted to characterize the use of specific gear and teeth features [E11]. Average and tendency indicators were simply extracted from the 10 Hz sampled contextual variables.

Modeling and diagnosis

After a huge preprocessing task, unitary predictive models were introduced to monitor specific energy pointer, as a target variable function of the indicators of the contextual variables. In the ADHER project, SVM with Gaussian Kernel were used to estimate the conditional density (function g) in a similar way that in the SPEC+ project. The final diagnosis was computed using an aggregation of the results provided by the 80 unitary predictive models implemented to monitor the equipment (rotating part of the helicopter). The diagnosis thresholds were computed using the available 10 sec records tagged with failure.

Software evaluation and testing

The diagnosis software was installed and evaluated by both companies: RSL and Eurocopter using real vibration data of a new helicopter fleet which were not included in the database used for self-learning software tools development. The test database included data from time periods corresponding to both normal and abnormal behavior of mechanical components. The "normal" and "abnormal" databases included respectively 350 flight hours from 7 helicopters : 230 flight hours of normal flights (880 vibration recordings) and 120 flight hours of abnormal flights (158 vibration recordings) corresponding to 3 types of failure. The testing results show that no Missed Detections were detected by the software and False Alarm Rate of 15 alarms per 1000 flight hours is estimated approximated results of 350 flight hours. These performances were qualified as very "good" results by the consortium [E11].

4.3 Health equipment monitoring with predictive modeling

Nowadays, a challenge for Health Usage Monitoring System (HUMS) is to implement automated low cost condition based maintenance systems as an alternative to equipment periodic inspections. Existing HUMS technologies, which generally propose in general basic diagnoses on data sensor and which simply rely on fixed alarm thresholds tend to generate high rate of false alarms. In this context, *predictive models* designed with contextual variables, may bring an efficient answer to the question of health monitoring.

4.3.1 Supervised classification vs anomaly detection

When the objective is to predict the health status of an industrial equipment, it may seem natural, at a first glance, to define a binary target variable to code the equipment status (0/1 for normal or abnormal work) and to train a classification model to explain the status of the equipment function of some co variables. These kinds of approaches, often used in Banks or Insurances companies for scoring, rely on the availability of historical databases, storing both the status and the explanatory variables for many observations. For health equipment monitoring, when the classification framework is chosen (and it is sometimes the case...), it quickly appears to be inappropriate. The status of the equipment is rarely available in the databases and if so, the alarm frequency corresponding to the abnormal working status is most of the time extremely low. In this case, classification models are inappropriate. Specific experimental designs are sometimes proposed in order to get supervised data from both status to calibrate classification method, but they often take a long time and are costly. Moreover, they are barely representative of usual operational work.

In this apparently unsupervised context, one needs to choose regression models, to monitor then diagnose the equipment. With an expert of the field, a target quantitative variable (Y) is provided which can be linked to the health of the equipment given the values of some contextual variables (X) which are recorded and which are known to influence the evolution of the target variable.

4.3.2 Expert knowledge vs knowledge extraction

In order to be robust, a predictive model needs to use the right "inputs". Integrating, manually, appropriate knowledge into the model after interviews or discussions with human experts may be extremely valuable. Combination of both data-driven models and expert knowledge are, to my humble opinion, the wiser option, and especially for applications providing a huge amount of data. For example, in the ADHER project, the available data were defined by vibration data at a high sampling rate. The monitoring of the energy at some specific frequencies was directly linked to well known failures and, in this case

expert knowledge was essential to extract the good features. Knowledge is also fundamental for driving decisions in the development of model. Expert knowledge should always be required in order to obtain information to provide relevant variables for the desired research objectives. *The best predictive models are fundamentally influenced by a modeler combining expert and context knowledge of the problem*. However, when it is not possible, automatic statistical investigation may supply.

At Miriad Technologies, Mutual information ratio (MIR) based on conditional entropy were computed to quantify the relationships between two random variables X and Y [Lacaille, 2003]: mutual information ratio is model-independent and can also be used, before modeling, to select the most relevant variables. In the multivariate framework, we developed a greedy algorithm based on MIR to select, based on the data, a small group of variables with a high mutual information ratio. It was used to mine the GIDAS database, one of the largest German In-Depth Accident Study [C5].

4.3.3 Robustness of an automatic decision in an operational environment

During a proof of concept, statistical models are built, calibrated and tested off line given usually well chosen historical data. When the models are running online, a key point is to be sure that the current models are well calibrated regarding the tasks of monitoring and diagnosis they have to provide. Consequently, one must be notified when the calibration of the models becomes obsolete. When a decision making software is used by people with statistical skills, they can easily detect an inappropriate model by themselves, and the task of updating a new model is usually performed manually. In an operational context, users are, most of the time, not statisticians, and the solution must guarantee by itself the use of an appropriate model. As it may be dangerous, for diagnosis purpose, to update the models automatically, it is at least possible to warn the operators that a model may be out of order and that the results provided by the monitoring software should not be taking temporarily taken into account. Without any such a feature, operational users will quickly give up a solution based on self learning algorithms.

4.4 Conclusions

Nowadays, information systems are systematically installed in industrial environment. Industrialization of statistical or machine learning algorithms becomes a key tool for decision making processes and in particular for health monitoring. The need for "data analysis" (in the large sense) and predictive modeling in private companies has increased in recent years. Mathematics, and in particularly statistics, appears as an essential asset for innovation and competitiveness [CMI and AMIES, 2015]. Consequently, many collaboration opportunities in this area have emerged for public laboratories. However, it is always a challenging task to successfully develop innovative scientific research and to respond, at the same time, to a real operational need. From my point of view, specific structures are still needed to make the bridge between the two worlds (the Academic Research *and] the Development* and to transform statistical innovative methods into useful operational software.

Recently, the way to handle Proof of concepts may have changed for large companies. Today, a POC may not be realized through a one to one partnership between an industrial and a research team but may be realized through open competitions proposed through different web sites (cf Kaggle or datascience.net). The codes, developed for the POC, are submitted to the competition web site and the corresponding performances are automatically computed. This let to work many teams in parallel on the same subject and to select at the end the best solutions.

Perspectives

A central objective of my work will remain to develop links between statistical or machine learning advances and industrial applications. To illustrate this general direction, I propose further developments of the work presented in Chapter 2, "From functional regression to electrical consumption forecast", that will be representative of my future projects.

- **Practical choice of a generic dictionary for a sparse representation.** Concerning the functional regression in high dimension, the assumptions of having a "sparse representation on a given dictionary of functions" is crucial. The choice of elaborating such a dictionary is however not straightforward. Up to now, the generic dictionary designed to produce a sparse representation of the functional signals has been elaborated manually. For the set of intra day load curves, the sparsity of different dictionaries has been studied and the Haar and Trigonometric dictionary has finally be chosen. Because of the low coherence assumption of the Learning Out of Leader algorithm, it is not possible at the present time to aggregate a large set of bases and to automatically select the best set of functions to obtain the "sparsest" representation. More generally, finding an algorithm able to automatically propose the best set of generic functions belonging to different basis in a reasonable computational time is a first challenging question.

- **Variable selection, feature extraction and grouping** A second research direction will be to study in the modeling context the benefits of an automatic variable representation. An important point would be to take into account the constraints of designing algorithms with a relative complexity to allow handling the requested large number of variables. For the modeling of the intra day load curves, a large set of meteorological variables were initially available (39 for temperature and cloud covering signals, and 293 for wind signals). These variables exhibited strong correlations due to spatial relations. In this context, up to this time, we chose to extract some basic features (min, max...), but more elaborate methods such as automatic grouping or model based clustering may be introduced to take into account the spatio-temporal relationships between these variables.

- **Probabilistic forecast.** A very challenging direction of research for improving the forecast of the intra day load curve, so far only provided by a simple curve, could be to compute probabilistic forecast using confidence intervals. Up to now, most of the confidence intervals are built point by point and, practically, bootstrap methods are used to provide envelopes around the curve. A collaboration with a team with theoretical expertise could lead to propose more functional confidence intervals around critical points of the curves.

Parts of these research directions will be developed within the framework of the ANR project FOREWER, in which I lead the task "From resource distribution to power consumption". Most of the statistical components of the FOREWER were directly designed based on the experience acquired in the RTE project.

Teaching

Regarding my teaching activity, a project close to my heart is to create a data mining course based on data acquired with connected objects carried by students such as accelerometer sensors. Various questions may be asked analyzing those data, as for example, the monitoring of the activity of a single subject or a comparison of the activity of a subject relative to the others. Different levels of analysis may be introduced using aggregated (daily) or raw data (signals). The project would aim to introduce the different data mining methodologies (regression, classification, clustering) and to illustrate their use by analyzing the data acquired by each student (or by the cohort of the class). One of the main technical difficulties would be to be able to extract the raw data from the connected objects. To overcome this difficulty, I propose to establish collaborations with the physics department in order to design the appropriate sensors during a preliminary joint work.

Industrialization of R&D

In the past years, I developed a strong experience in the use of machine learning or statistics to industrialize a solution. To undertake that kind of project, you need to be successively a salesman to quantify the costs, a lawyer to set up the contract, a researcher to design the solution, a computer scientist to code the program, and a support service to answer the potential questions of the users. Moreover, complementary developments are often necessary to publish the method and the results. Of course, nobody cumulates all these expertises but still, it is necessary to develop at least some basic knowledge in all these fields.

Nowadays, various structures emerge that aim to facilitate public-private collaborations. For example, the AMIES agency promotes meetings between industrials and researchers to favor collaborations. The "Société d'Activation et de Transfert Technologique" (SATT) helps for the valorization of the contracts. Nevertheless, beyond organizing the first contact, a lot could be done to support the research teams along the whole collaboration process. This is particularly challenging in the field of mathematics, which accounts today for 15% of the French Gross Domestic Product [CMI and AMIES, 2015]. Based on my past experience in R&D, I could play an active role to contribute to this objective.

Bibliography

A. Antoniadis, E. Paparoditis, and T. Sapatinas. A functional wavelet–kernel approach for time series prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(5):837–857, 2006.

S. Arlot, A. Celisse, et al. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79, 2010.

F. R. Bach. Consistency of the group lasso and multiple kernel learning. *The Journal of Machine Learning Research*, 9:1179–1225, 2008.

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

K. Bertin, E. Le Pennec, and V. Rivoirard. Adaptive dantzig density estimation. In *Annales IHP, Probabilités et Statistiques*, volume 47, pages 43–74, 2011.

P. Besse, C. Le Gall, N. Raimbault, and S. Sarpy. Data mining et statistique. *Journal de la société française de statistique*, 142(1):5–36, 2001.

P. J. Bickel, J. B. Brown, H. Huang, and Q. Li. An overview of recent developments in genomics and associated statistical methods. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4313–4337, 2009a.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009b.

L. Birgé and P. Massart. Minimal penalties for gaussian model selection. *Probability theory and related fields*, 138(1-2):33–73, 2007.

M. Bourriga and V. Lefieux. Etude du modele de consommation electrique. Technical report, RTE, 2014.

L. Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001.

L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and regression trees*. CRC press, 1984.

P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer, 2011.

P. Bühlmann, P. Rütimann, S. van de Geer, and C.-H. Zhang. Correlated variables in regression: clustering and sparse estimation. *Journal of Statistical Planning and Inference*, 143(11):1835–1858, 2013.

E. Candes and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

E. J. Candès and Y. Plan. Near-ideal model selection by l1 minimization. *The Annals of Statistics*, 37(5A): 2145–2177, 2009.

E. J. Candes and T. Tao. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51 (12):4203–4215, 2005.

O. Catoni. Statistical learning theory and stochastic optimization, lectures on probability theory and statistics, saint-flour xxxi–2001, volume 1851 of lecture notes in mathematics. *Lecture Notes in Mathematics*, pages 1–269, 2004.

Y. Chakhchoukh, P. Panciatici, and P. Bondon. Robust estimation of sarima models: Application to short-term load forecasting. In *Statistical Signal Processing, 2009. SSP'09. IEEE/SP 15th Workshop on*, pages 77–80. IEEE, 2009.

H. Cho, Y. Goude, X. Brossat, and Q. Yao. Modeling and forecasting daily electricity load curves: a hybrid approach. *Journal of the American Statistical Association*, 108(501):7–21, 2013.

W. Christiaanse. Short-term load forecasting using general exponential smoothing. *Power Apparatus and Systems, IEEE Transactions on*, (2):900–911, 1971.

CMI and AMIES. A study of the socio-economical impact of mathematics in france. Technical report, CMI, 2015.

J. Cugliari. *Prévision non paramétrique de processus à valeurs fonctionnelles: application à la consommation d'électricité*. PhD thesis, Université Paris Sud-Paris XI, 2011.

A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In *Learning theory*, pages 97–111. Springer, 2007.

M. Devaine, P. Gaillard, Y. Goude, and G. Stoltz. Forecasting electricity consumption by aggregating specialized experts. *Machine learning*, 90(2):231–260, 2013.

E. Devijver. Model-based clustering for high-dimensional data. application to functional data. *arXiv preprint arXiv:1409.1333*, 2014.

D. Donoho and J. Jin. Higher criticism thresholding: optimal feature selection when useful features are rare and weak. *Proceedings of the National Academy of Sciences*, 105(39):14790–14795, 2008.

D. L. Donoho. For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. *Communications on pure and applied mathematics*, 59(6):797–829, 2006.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2): 407–499, 2004.

B. Efron, T. Hastier, and R. Tibshirani. Discussion: The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2358–2364, 2007.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*, 2010.

P. Gaillard and Y. Goude. Forecasting the electricity consumption by aggregating experts; how to design a good set of experts. 2014.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

M. T. Hagan and S. M. Behr. The time series approach to short term load forecasting. *Power Systems, IEEE Transactions on*, 2(3):785–791, 1987.

T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani. *The elements of statistical learning*, volume 2. Springer, 2009.

J. Huang, T. Zhang, et al. The benefit of group sparsity. *The Annals of Statistics*, 38(4):1978–2004, 2010.

J. Huang, P. Breheny, and S. Ma. A selective review of group selection in high-dimensional models. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 27(4), 2012.

R. J. Hyndman and S. Fan. Density forecasting for long-term peak electricity demand. *Power Systems, IEEE Transactions on*, 25(2):1142–1153, 2010.

A. Juditsky and A. Nemirovski. Functional aggregation for nonparametric regression. *Annals of Statistics*, pages 681–712, 2000.

R. Klein. Method and system for diagnostics and prognostics of a mechanical system, Apr. 11 2006. US Patent 7,027,953.

J. Lacaille. *Industrialisation d'algorithmes mathématiques*. PhD thesis, Université Paris 1, 2003.

K. Lounici, M. Pontil, S. Van De Geer, A. B. Tsybakov, et al. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39(4):2164–2204, 2011.

J. Mairal and B. Yu. Grouping strategies and thresholding for high dimensional linear models: Discussion. *Journal of Statistical Planning and Inference*, 143(9):1451–1453, 2013.

F. Marin, F. Garcia-Lagos, G. Joya, and F. Sandoval. Global model for short-term load forecasting using artificial neural networks. *IEE Proceedings-Generation, Transmission and Distribution*, 149(2):121–125, 2002.

P. Massart, C. Meynet, et al. The lasso as an l1-ball model selection procedure. *Electronic Journal of Statistics*, 5:669–687, 2011.

N. Meinshausen. Grouping strategies and thresholding for high dimensional linear models: Discussion. *Journal of Statistical Planning and Inference*, 143(9):1439–1440, 2013.

N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, pages 246–270, 2009.

M. Mougeot. Détection automatique sous mdscan des fluctuations de tension et des variations thd sur signaux électriques, schneider industrie. Technical report, Miriad Technologies, 2000.

M. Mougeot. Composant auto-adaptatif pour le suivi des performances énergétiques de compresseurs, air liquide. Technical report, Miriad Technologies, 2003.

M. Mougeot. Données ferroviaires de singapour. synchronisation des données et détection automatique du sens de manoeuvre. siema applications. Technical report, Miriad Technologies, 2005.

M. Mougeot and Fakhr-eddine. Méthodologie de diagnostics de cordons de soudure, air liquide, centre technique des applications de soudure (ctas). Technical report, Miriad Technologies, 2002.

M. Mougeot and L. Layeillon. Aide à la conduite de réacteurs catalytiques de production d avm. Technical report, Miriad Technologies, 2004.

M. Mougeot and A. Maazi. Amélioration de la maîtrise de la production de latex, rhodia industrie. Technical report, Miriad Technologies, 2000.

M. Mougeot, O. Gérard, and R. Azencott. Product quality stabilization (c2mnp) on rhodia site (winder/usa). Technical report, Miriad Technologies, 2003.

D. Needell and J. A. Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.

G. Obozinski. Grouping strategies and thresholding for high dimensional linear models: Discussion. *Journal of Statistical Planning and Inference*, 143(9):1441–1446, 2013.

J.-M. Poggi. Prévision non paramétrique de la consommation électrique. *Revue de Statistique Appliquée*, 42(4):83–98, 1994.

J. W. Taylor. Triple seasonal methods for short-term electricity demand forecasting. *European Journal of Operational Research*, 204(1):139–152, 2010.

J. W. Taylor. Short-term load forecasting with exponentially weighted methods. *Power Systems, IEEE Transactions on*, 27(1):458–464, 2012.

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

J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *Information Theory, IEEE Transactions on*, 53(12):4655–4666, 2007.

A. B. Tsybakov. Optimal rates of aggregation. In *Learning Theory and Kernel Machines*, pages 303–313. Springer, 2003.

S. van de Geer. Grouping strategies and thresholding for high dimensional linear models: Discussion. *Journal of Statistical Planning and Inference*, 143(9):1447–1450, 2013.

S. A. Van De Geer, P. Bühlmann, et al. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

V. N. Vapnik. Estimation of dependences based on empirical data. *NY: Springer-Verlag*, 1982.

H. Wickham. Data science: how it is different to statistics? *IMS Bulletin*, 43(6):7, 2014.

M. Yuan. Grouping strategies and thresholding for high dimensional linear models: Discussion. *Journal of Statistical Planning and Inference*, 143(9):1454–1456, 2013.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, pages 3468–3497, 2009.

Part III

Annexes

Industrial technical reports

1. Mougeot M., Cherif O. (2005) SPEC+ sofware for Air Liquid America. Technical document. 15/08/2005
2. Mougeot M. (2004) Données ferroviaires de Singapour. Synchronisation des données et détection automatique du sens de manoeuvre. p 1-24. Rapport technique confidentiel Siema Applications.
3. Mougeot M., Layeillon L. (2004) Aide à la conduite de réacteurs catalytiques de production d AVM. Rapport technique, confidentiel industrie. 21/06/2004.
4. Mougeot M. (2003) Composant auto-adaptatif pour le Suivi des Performances Energétiques de Compresseurs, Cahier des charges Installation de SPEC+ sur site Pilote. Rapport technique. Doc. confidentiel Industrie Air Liquide.
5. Mougeot M., Gérard O., Azencott R. (2003) Product quality stabilization (C2MNP) on Rhodia site (Winder/USA). Synthetic report, p 1-29. Confidential report Rhodia/HPCII . V1.0 April 3rd 2003.
6. Mougeot M., Fakhr-eddine K. (2002) Méthodologie de diagnostics de cordons de soudure. P 1-30. Rapport technique confidentiel Air liquide, Centre Technique des Applications de Soudure (CTAS). V1.0 février 2002.
7. Mougeot M., Gerard O. (2002) Influence Analysis for the Miranol C2MNP product. Confidential technical report Rhodia/HPCII. 1-15. V1.0 December 2002.
8. Mougeot M. (2001) Diagnostics virtuels de qualité pour le procédé de fabrication du vaccin contre la coqueluche. p 1-30. Rapport technique confidentiel Rhône-Poulenc, Décembre 2001.
9. Mougeot M., Besnard J. (2000) Détection automatique sous MdScan des fluctuations de tension et des variations THD sur signaux électriques. p 1-28. Rapport technique confidentiel Industrie Schneider Industrie. Septembre 2000.
10. Mougeot M. (2000) Synthèse des travaux du projet PACTE. Rapport technique Miriad Tech.. Décembre 2000.
11. Mougeot M., Gaudier F. (2000) Quantification automatique de l' impact de l' analyse d images sur la qualité et la productivité de fermentations enzymatiques. P 1-29. Projet PACTE. Rapport Tec Lesaffre industrie. V1.0
12. Maazi A., Mougeot M. (2000) Amélioration de la maîtrise de la production de Latex, p 1-27. Rapport confidentil Rhodia Industrie. V1.0 Décembre 2000.
13. Mougeot M. (1999) Analyse d'influence sur la qualité des poudres compactées Christian Dior. P 1-31. projet PACTE, rapport technique confidentiel industrie. V1.0 10 Juin 1999.
14. Mougeot M. (1999) Analyse des facteurs influents sur la qualité des comprimés pharmaceutiques, p 1-2. projet PACTE, Rapport technique confidentiel Rhône-Poulenc, V1.0 Octobre 1999.

15. Azencott R., Mougeot M. (1995) Détection d'anomalies par méthodes neuronales : applicationà la réaction de désaminiation du DCNA. Rapport technique Rhône-Poulenc Miriad 1995.

16. Azencott R., Catoni O. , Mougeot M. (1994) Diagnostic neuronal Multicapteurs: application au démarrage du moteur Vulcain. Rapport technique CNES Miriad 1994.