LABEX-ECODEC
Axis 5: New Challenges for New Data
Activity report, October 2014

1 Main research areas

In the recent years the quantity and quality of data available has exponentially increased. Sources of data are more and more diverse, and each source generates vast amounts of information. For example, the internet generates data on social networks, firms have access to data with more variables than observations, administrative data are more and more rich and may include the entire population (on employment, health…). The quantities produced are so enormous that classical statistical methods are often not adapted to handle them, either because of the number of variables or because of the number of observations.
Axe 5 intends to develop research in the following directions:

- Development of new statistical tools for conducting inference in presence of this new type of data and study of their theoretical properties. A particular emphasis is addressed to the study of high dimensional models with sparsity and of new Markov Chain Monte (MCMC) algorithms.
- Use of these new statistical tools to develop empirical analysis of: network models in economics and marketing, models with social interactions and consumption models. In addition, a particular focus will be put on the question of identification of these models.
- Study of other applications linked to “Big Data”.

2 Composition of the team

The axis “New Challenges for New Data” is directed by A. Simoni (Principal Investigator, ENSAE-CREST – replacement of E. Gautier in Sept. 2014), G. Stoltz (co-Principal Investigator, HEC Paris – replacement of J. Hombert in Sept. 2014) and E. Strobl (co-Principal Investigator, Ecole Polytechnique).

Our team is composed of economists, applied mathematicians and statisticians of Ecole Polytechnique, HEC Paris and ENSAE-ParisTech.

Researcher		Institution
ALQUIER	Pierre	CREST-ENSAE
BELZIL	Christian	ECOLE POLYTECHNIQUE
CALVET	Laurent	HEC PARIS
CHOPIN	Nicolas	CREST-ENSAE
COTTET	Vincent	CREST-ENSAE
DALALYAN	Arnak	CREST-ENSAE
D'HAULTFOEUILLE	Xavier	CREST-ENSAE
EBBES	Peter	HEC PARIS
FERMANIAN	Jean-David	CREST-ENSAE
HOMBERT	Johan	HEC PARIS
KRAMARZ	Francis	CREST-ENSAE
LOPEZ	Olivier	CREST-ENSAE
MEJEAN	Isabelle	ECOLE POLYTECHNIQUE
PATNAM	Manasa	CREST-ENSAE
ROUSSEAU	Judith	CREST-ENSAE
SIMONI	Anna	CNRS and CREST-ENSAE
SOTGIU	Francesca	HEC PARIS
STOLTZ	Gilles	HEC PARIS
STROBL	Eric	ECOLE POLYTECHNIQUE
TSYBAKOV	Alexandre	CREST-ENSAE
YANG	Cathy Liu	HEC PARIS

Participation in Editorial Boards:

- *Annals of Statistics* (J. Rousseau, A. Tsybakov)
- *Annales d'Economie et Statistique* (F. Bloch, X. D'Haultfoeuille, F. Kramarz)
- *Asian Journal of Management Science and Applications* (B. Huang)
- *Australian and New-Zealand Journal of Statistics* (J. Rousseau)
- *Bernoulli* (J. Rousseau, A. Tsybakov)
- *Economics Letters* (F. Bloch)
- *Economie Internationale* (E. Strobl)
- *Electronic Journal of Statistics* (A. Dalalyan, A.Tsybakov)
- *Games and Economic Behavior* (F. Bloch)
- *Journal of Fractal Geometry* (L. Calvet)
- *Journal of the Japan Statistical Society* (A.Dalalyan)

- *Journal of Population Economics* (F. Kramarz)
- *Journal of Public Economic Theory* (F. Bloch)
- *Journal of the Royal Statistical Society, ser. B* (N.Chopin)
- *Journal of Statistical Planning and Inference* (A.Dalalyan, A.Tsybakov)
- *Labor Economics* (F. Kramarz)
- *Mathematical Social Sciences* (F. Bloch)
- *Research in Economics* (F. Kramarz)
- *Recherches Economiques de Louvain* (F. Kramarz)
- *Statistical Methods and Applications* (N. Chopin)
- *Statistical Inference for Stochastic Processes (*A.Dalalyan*)*
- *Statistics and Computing* (N. Chopin)

Distinctions and Prizes received:

- F. Kramarz : elected fellow of the *Econometric Society* (2013)
- A. Tsybakov : Humboldt prize (2013)
- F. Kramarz : Keynote speaker at CAED, Nürnberg (2012).
- F. Kramarz : Keynote speaker at Revue Concurrences (2013).
- F. Kramarz : Keynote speaker at 25th EALE conference (2013).
- F. Kramarz : Keynote speaker at Zürich workshop on Economics, (2013)
- A. Tsybakov : lecture "Aggregation and high-dimensional statistics"at the 43rd Saint-Flour Probability Summer School (France).
- A. Tsybakov : 2012 IMS Medallion Lecture
- A. Tsybakov : invited Lecture at 2014 International Congress of Mathematicians
- A. Dalalyan : "Notable paper award" at the conference AI-STATS (2013)
- E. Gautier : ERC starting grant (2013)
- E.Gautier (coPI), J. Rousseau, A. Simoni (PI), A. Tsybakov : ANR Blanc IPANEMA grant for the period 2013-2017
- F. Kramarz: AXA research fund

3 Publications

In Subsections 1 and 2 we display a selection of published papers and working papers produced by the team within the axis topics. The papers In Subsections 3 and 4 have been specifically funded by the Labex Ecodec.

1) Published and accepted papers

- Gautier, E., Tsybakov, A.B. (2013). Pivotal estimation in high-dimensional regression via linear programming. In: *Empirical Inference -- Festschrift in Honor of Vladimir N. Vapnik*, B.Schölkopf, Z. Luo, V. Vovk eds., 195 - 204. Springer, New York e.a.
- Rosenbaum, M., Tsybakov, A.B. (2013). Improved matrix uncertainty selector. In: *From Probability to Statistics and Back: High-Dimensional Models and Processes -- A Festschrift in Honor of Jon A. Wellner*, M.Banerjee et al. eds. IMS Collections, v.9, 276-290. Institute of Mathematical Statistics.
- Tsybakov, B.S., Tsybakov, A.B. (2012). On Walsh code assignment. *Problems of Information Transmission*, v.48, 334-341.
- Rigollet, P., Tsybakov, A.B. (2012). Estimation of covariance matrices under sparsity constraints. *Statistica Sinica*, v.22, 1358-1365.
- Dalalyan, A., Tsybakov, A.B. (2012). Sparse Regression Learning by Aggregation and Langevin Monte-Carlo. *Journal of Computer and System Sciences*, v.78, 1423-1443.
- Dalalyan, A., Tsybakov, A.B. (2012). Mirror averaging with sparsity priors. *Bernoulli*, v.18, 914-944.
- Giraud, C., Tsybakov, A.B. (2012). Discussion of "Latent variable graphical model selection via convex optimization". *Annals of Statistics*, v.40, 1984-1988.
- Rigollet, P., Tsybakov, A.B. (2012). Sparse estimation by exponential weighting. *Statistical Science*, v. 27, 558–575.
- L. Comminges, A.S. Dalalyan (2013). Minimax testing of a composite null hypothesis defined via a quadratic functional in the model of regression. *Electronic Journal of Statistics*, 7, pp. 146–190.
- A.S. Dalalyan, M. Hebiri, K. Meziani, J. Salmon (2013). Learning Heteroscedastic Models by Convex Programming under Group Sparsity. In *Journal of Machine Learning Research* - W & CP 28(3) (ICML 2013), pp. 379–387.
- O. Collier, A.S. Dalalyan (2013). Permutation estimation and minimax rates of identifiability. In *Journal of Machine Learning Research* - W & CP 31 (AI-STATS 2013), pp. 10-19.

- J. Courchay, A.S. Dalalyan, R. Keriven, P.Sturm (2012). On Camera calibration with linear programming and loop constraint linearization. *Int. J. Comput. Vis.*, 97(1), pp. 71–90.
- A.S. Dalalyan, R. Keriven (2012). Robust estimation for an inverse problem arising in multiview geometry. *J. Math. Imaging Vision*, 43(1), pp. 10–23.
- A.S. Dalalyan (2012). SOCP based variance free Dantzig Selector with application to robust estimation. *C. R. Math. Acad. Sci. Paris*, 350(15-16), pp. 785–788.
- A.S. Dalalyan, Joseph Salmon (2012). Sharp Oracle Inequalities for Aggregation of Affine Estimators. *Ann. Statist*., 40(4), pp. 2327–2355.
- L. Comminges, A.S. Dalalyan (2012). Tight conditions for consistency of variable selection in the context of high dimensionality. *Ann. Statist*, 40(5), pp. 2667-2696.
- A.S. Dalalyan and O. Collier (2012). Wilks' phenomenon and penalized likelihood-ratio test for nonparametric curve registration. In *Journal of Machine Learning Research - Proceedings Track 22*, pp. 264-272.
- A.S. Dalalyan, Y. Chen (2012). Fused sparsity and robust estimation for linear models with unknown variance. In *Advances in Neural Information Processing Systems 25*: NIPS, pp. 1268-1276.
- S. Petrone, J. Rousseau and C. Scriocciolo (2013). Bayes and empirical Bayes : Do they merge? To appear in *Biometrika.*
- J.M. Marin, N. Pillai, C.P. Robert and J. Rousseau (2013). Relevant statistics for Bayesian model choice. To appear in *JRSS B*.
- E. Gassiat and J. Rousseau (2013). On the asymptotic behaviour of the posterior distribution in hidden Markov Models. To appear in *Bernoulli*.
- J. Arbel, G. Gayraud and J. Rousseau (2013). Bayesian optimal estimation using a sieve prior. *Scandinavian J. Statist.* 40, 549–570
- V. Rivoirard and J. Rousseau (2012). Bernstein-von Mises Theorem for linear functionals of the density, *Annals of Statistics*, 40, 1489-152
- McVinish, R. Mengersen, K., Rousseau J., Nur, D. and Guihenneuc C. (2012). Recentered importance sampling with applications to Bayesian model validation. *JCGS*.
- E. Gautier and Y. Kitamura (2013). Nonparametric estimation in random coefficients binary choice models, *Econometrica* 81, 581-607.
- M. Patnam and C. Helemers, Does the Rotten Child Spoil His Companion? Spatial Peer Effects Among Children in Rural India, forthcoming, *Quantitative Economics.*
- M. Patnam and P. Krishnan Neighbours and Extension Agents in Ethiopia: Who matters more for technology diffusion?, forthcoming, *American Journal of Agricultural Economics*.

- F. Kramarz, O. Nordström Skans (2013). When Strong Ties are Strong: Networks and Youth Labor Market Entry. *The Review of Economic Studies*, forthcoming.
- J. Abowd, F. Kramarz, P. Lengermann, K. Mc Kinney, S. Roux (2013). Persistent Inter-Industry Wage Differences: Rent-Sharing and Opportunity Costs. I*ZA Journal of Labor Economics*, forthcoming.
- F. Kramarz, D. Thesmar (2013). Networks in the Boardroom. *Journal of the European Economic Association*, 11(4), 780-807.
- Florens, J-P. and A. Simoni (2012). Nonparametric Estimation of an Instrumental Regression: a Quasi-Bayesian Approach Based on Regularized Posterior. *Journal of Econometrics*, 170(2), p.458-475.
- Florens, J-P. and A. Simoni (2012). Regularized Posteriors in Linear Ill-Posed Inverse Problems". *Scandinavian Journal of Statistics*, Vol.39(2), 214-235.
- Ebbes P., J. Liechty and R. Grewal. Attribute-Level Heterogeneity. *Management Science*,
- Ebbes P., W. DeSarbo, D. K. H. Fong (2012). A hierarchical Bayesian regression model for cross sectional data involving a single observation per response unit, *Psychometrika*, 2012, vol. 77, pp. 293-314.
- Paris, Q. (2014). Minimax adaptive dimension reduction for regression. *Journal of Multivariate Analysis,* forthcoming.

2) Working papers and papers in revision

- E. Gautier and S. Hoderlein (2013). A triangular treatment effect model with random coefficients in the selection equation.
- E. Gautier and E. Le Pennec (2013). Adaptive estimation in the nonparametric random coefficients binary choice model by needlet thresholding
- E. Gautier and A.B. Tsybakov (2013). High-dimensional instrumental variables regression and confidence sets, revise and resubmit for *Econometrica.*
- Liao, Y. and A. Simoni (2013). Semi-parametric Bayesian Partially Identified Models based on Support Function. arXiv:1212.3267.

3) Labex Ecodec published and accepted papers

- Tsybakov, A.B. (2014). Aggregation and minimax optimality in high-dimensional estimation. In: *Proceedings of the International Congress of Mathematicians*, Seoul, August 2014, to appear.
- Kerkyacharian, G., Tsybakov, A.B., Temlyakov, V., Picard, D., Koltchinskii, V. (2014). Optimal exponential bounds on the accuracy of classification. *Constructive Approximation*, v.39, 421-444.
- Dalalyan, A., Ingster, Y., Tsybakov, A.B. (2014). Statistical inference in compound functional models. *Probability Theory and Related Fields,* v.158, n.3-4, 513-532.
- A.S. Dalalyan and O. Collier (2012). Wilks' phenomenon and penalized likelihood-ratio test for nonparametric curve registration. In *Journal of Machine Learning Research - Proceedings Track 22*, pp. 264-272.
- S. Barthelmé and N. Chopin (2014). Expectation-propagation for likelihood-free inference. *Journal of the American Statistical Association*, 109(505):315-333.
- N. Chopin and S.S. Singh (2014). On the particle Gibbs sampler. *Bernoulli* (in press).
- N. Chopin, P. Jacob, and O. Papaspiliopoulos (2013). SMC2: A sequential Monte Carlo algorithm with particle Markov chain Monte Carlo updates. *Journal of the Royal Statistical Society (series B)*, 75(3):397-426.
- N. Chopin, J. Rousseau, and B. Liseo (2013). Computational aspects of Bayesian spectral density estimation. *Journal of Computational and Graphical Statistics*, 22(3):533-557.
- S. Singh, N. Whiteley, and N. Chopin (2013). Bayesian learning of noisy Markov decision processes. *ACM Transactions on Modeling and Computer Simulation* (TOMACS), 23(1).
- N. Chopin, T. Lelièvre, and G. Stoltz (2012). Free energy methods for Bayesian inference: Efficient exploration of univariate gaussian mixture posteriors. *Statistics and Computing*, (4):897-916.
- C. Schäfer and N. Chopin (2012). Adaptive Monte Carlo on binary sampling spaces. *Statistics and Computing*, 23(2):163-184.

4) Ecodec Working papers

- Belloni, A., Rosenbaum, M., Tsybakov, A.B. (2014). Linear and convex programming estimators in high-dimensional errors-in-variables models. arxiv1408.0241
- Rakhlin, A., Sridharan, K., Tsybakov, A.B. (2014). Empirical entropy, minimax regret and minimax risk.

- A.S. Dalalyan, M. Heiri, J. Lederer (2014). On the Prediction Performance of the Lasso. *arXiv. 1402.1700*.
- O. Collier, A.S. Dalalyan (2013). Minimax rates in permutation estimation for feature matching. *arXiv. 1310.4661*.
- S. Barthelmé and N. Chopin (2014). The Poisson transform for unnormalised statistical models. *ArXiv preprint, (1406.2839)*.
- P. Alquier, V. Cottet, N. Chopin, and J. Rousseau (2014). Bayesian matrix completion: prior specification and consistency. *ArXiv preprint*, (1406.1440).
- M. Gerber and N. Chopin (2014). Sequential Quasi-Monte Carlo. *ArXiv preprint*, (1402.4039).
- E. Strobl, Natural Disasters and the Death, Life, and Birth of Plants: The Case of the Kobe Earthquake.
- Ebbes, P., O. Netzer. Using hidden Markov models to identify job seekers from social network data.
- Belloni, A., Rosenbaum, M., Tsybakov, A.B. (2014). Linear and convex programming estimators in high-dimensional errors-in-variables models. arxiv1408.0241
- Rakhlin, A., Sridharan, K., Tsybakov, A.B. (2014). Empirical entropy, minimax regret and minimax risk.
- A.S. Dalalyan and O. Collier (2013). Curve registration by nonparametric goodness-of-fit testing. hal-00580047.
- Belzil C., Hansen, J. and X. Liu (2014). Dynamic Skill Accumulation, Education Policies and the Return to Schooling, Revised and resubmitted at *Quantitative Economics.*
- Belzil C., Hansen, J. and X. Liu (2014). Estimation of a Dynamic Skill Accumulation Model with Comparative Advantages and its Economic Implications, Revised and Resubmitted at *Journal of Political Economy.*
- Amat C., T. Michalski and G. Stoltz (2014). Forecasting exchange rates better than the random walk thanks to machine learning techniques.
- M. Hoffman, J. Rousseau and Johannes Schmidt-Hieber (2013): On adaptive posterior concentration rates. *Revision in the Annals of statistics*
- I. Castillo and J. Rousseau (2013). A general Bernstein von Mises Theorem in semi-parametric models. *Revision in the Annals of statistics*
- S.Donnet, V.Rivoirard, J.Rousseau and C.Scricciolo (2014) Posterior concentration rates for Aalen counting processes.
- S.Donnet, V.Rivoirard, J.Rousseauand C.Scricciolo (2014) Posterior concentration rates for empirical Bayes procedures, with applications to Dirichlet Process mixtures.

4 Grants allocated to research projects

- F. Bloch (and PhD): Rumors on social networks : theory and empirics
- V-E. Brunel : Estimation de polytopes convexes
- N. Chopin (and PhD): Bayesian Computation 2.0
- A. Dalalyan: Estimation parcimonieuse, prédiction pour les séries chronologiques, réduction de dimension dans des modèles non-linéaires et détection de liens dans un graphe
- P. Ebbes: Client value in the presence of credit risk for Wehkamp.nl
- P. Ebbes : Predicting job seekers at LinkedIn using social network and social network activity data
- E. Gautier : High-dimension and random coefficients in econometrics
- M. Patnam : Observational learning in internet markets
- J. Rousseau (and PhD): Bayesian Nonparametric
- E. Strobl : The impact of the kobe earthquake on manufacturing : analysis using plant level data and damage map
- A. Tsybakov : Estimation and prediction in models described by high-dimensional matrices

Project descriptions

Project title: Rumors on Social Networks : Theory and Empirics
Principal investigator: Francis Bloch
(formerly Ecole Polytechnique, now University de Paris 1)

Due to Francis Bloch's departure from the department of economics in September 2013, the project is no longer financed by the LABEX.
Use of funds: The funds were used to finance 6 conference participations / research visits by Francis Bloch, as well as four conference participations by two of Francis Bloch's PhD students. Additionally, some funds were used to finance seminars given by external speakers in Paris, and other speakers.

Project title: Bayesian Computation 2.0

Principal Investigator: Nicolas Chopin (CREST-ENSAE)
Other co-authors: Adam Johansen (Warwick University), Harvard Rue (Trondheim University), Omiros Papaspiliopoulos (Pompeu Fabra), Jean-Michel Marin (Montpellier), Simon Barthelmé (Lausanne), Mathieu Gerber (Havard)

* "Sequential Quasi-Monte Carlo" (with M. Gerber, now in Harvard): This is certainly the most important and inpactful piece of work of this project. It is currently in revision as a read paper at the RSS (Royal Statistical Society). In that paper, we propose a variant of SMC (Sequential Monte Carlo) which converges much faster, thanks to the introduction of QMC (Quasi-Monte Carlo). This paper is very technical (as the introduction of QMC is completely new in SMC), and the proposed methodology improves on standard SMC by several orders of magnitude. Impacted applications include all standard applications of SMC, that is sequential analysis of hidden Markov models in signal processing, finance, neuro-sciences, and so on.

* « The Poisson transform for unnormalised statistical models » (with S. barthelmé, Lausanne) : What can a statistician do with an unnormalised likelihood ? ie. A likelihood defined up to an intractable normalising constant. In this paper, we develop an approach for this problem, based on a Poisson process representation of the generating data mechanism : in the end, one obtains a reformulated model where the normalising constant is simply an extra parameter to estimate. Applications include human vision (eye-tracking data), Ising models (for e.g. image processing), and so on.

* " On the particle Gibbs sampler" (with S.S. Singh, Cambridge, accepted in Bernoulli). This paper contains the first theoretical results on the PG (Particle Gibbs) algorithm, an efficient but technically challenging branch of PMCMC (Particle MCMC).Interest in this field is evidenced by the several papers submitted in the following year (but again we were first!).

*"Expectation-propagation for likelihood-free inference" (avec S. Barthelmé, published in JASA). This is the first paper that manages to adapt EP (Expectation Propagation, an approximation technique from machine learning) to ABC (Approximate Bayesian Computation). In this way, one manages to tackle intractable likelihoods, without resorting to “summary statistics” (as in standard ABC) which introduce a bias that is hard to quantify.

Project title: Sparse prediction in high-dimensional statistical models in order to detect links in graphs
Principal investigator: Arnak Dalalyan (CREST-ENSAE)

During the period under consideration, our attention focused on the problems of sparse prediction in high-dimensional statistical models and of detecting links in the graphs. In these problems significant progress has been achieved. The corresponding findings are reported in the following preprints and publications:

- Olivier Collier, Arnak S. Dalalyan (2013). Minimax rates in permutation estimation for feature matching. ArXiv. 1310.4661.

- Olivier Collier, Arnak S. Dalalyan (2013). Permutation estimation and minimax rates of identifiability. In Journal of Machine Learning Research - W & CP 31 (AI-STATS 2013), pp. 10-19.

- Arnak S. Dalalyan, Mohamed Hebiri, Katia Meziani, Joseph Salmon (2013). Learning Heteroscedastic Models by Convex Programming under Group Sparsity. In Journal of Machine Learning Research - W & CP 28(3) (ICML 2013), pp. 379-387.

- Arnak S. Dalalyan, Mohamed Heiri, Johannes Lederer (2014). On the Prediction Performance of the Lasso. arXiv. 1402.1700.

The first two papers concern the problem of link detection in bipartite graphs that naturally appears in the context of feature matching in digital images. The nodes of the corresponding graph represent the keypoints in each image,that can be grouped according to the image which they belong to. In the case of two images this leads to a bipartite graph: a link in the graph indicates that the endpoints of the link correspond to the projections of a same 3D point into the planes of the two images. This approach to link prediction is used in the pipeline of 3D reconstruction from 2D images. Our main contribution consists in introducing a new, global method of link prediction that is proved, both theoretically and empirically, to outperform the previously used greedy methods. This improvement is particularly significant when the features associated to each node are contaminated with a heteroscedastic noise with relatively large variance. The second paper received the notable paper award at the 16th International Conference on Artificial Intelligence and Statistics.

The last two papers in the foregoing list concern different aspects of the problem of sparse prediction:

- the possibility to extend the Lasso and L1-penalty based techniques to the case of heteroscedastic noise
- the influence of the correlations between the covariates on the prediction performance of the Lasso.

In both directions we obtained nearly exhaustive answers to the questions under consideration. The procedure providing an extension of the Lasso to the heteroscedastic noise was implemented under Matlab and the codes were made freely available.
Anteresting and unexpected outcome of the second study was the elaboration of a new technique of proof allowing to check the well-known (but hard to verify) assumptions on the design matrix in a multiple linear regression that imply strong theoretical guarantees both for the Lasso and for the Dantzig selector.

The funding offered by the Labex allowed us to attend several workshops and conferences that was extremely helpful in completing the research programme summarized above. The list of these conferences and workshops is :

- 2nd Congress of the International Society of Nonparametric Statistics, Spain, June 13, 2014
- Workshop "Adaptive Statistical Inference", Oberwolfach, Germany, March 14, 2014
- LDHD : Opening Workshop : September 8-12, 2013, SAMSI, USA
- Bernoulli Society Satellite Meeting to the ISI World Statist. Congress 2013, Tokyo, 4 Sept.
- International Conference "Mathematics in Armenia : Advances and Perspectives", Aug. 26
- European Meeting of Statisticians, Budapest, July 22, 2013
- International Workshop on Statistical Learning, Moscow, June 27, 2013
- SAPS IX workshop, Le Mans, France, March 15, 2013
- Workshop "Optimisation pour l'apprentissage statistique", Les Houches, Jan. 6, 2013

Project 1: Client value in the presence of credit risk for Wehkamp.nl
Project 2: Predicting job seekers at LinkedIn using social network and social network activity data.
Principal investigator: Peter Ebbes (HEC Paris)
Co-authors: Project 1: Jaap Wieringa (University of Groningen)
Project 2: Oded Netzer (Columbia University)

Project 1: we have made good progress on the project. In a nutshell, we will be looking at customer lifetime value in the presence of credit risk for a large European webshop. Many customers of the webshop purchase products on credit. However, customer relationship managers, who traditionally have been part of the marketing team, have neither access nor information regarding the financial behavior of the customer. Vice versa, the financial department has no information on the "client value" (i.e. has the client been loyal to the firm, bought a lot of products in the past etc). The question then becomes is whether both departments can make better decisions regarding the client (e.g. take more financial risk on a loyal client versus a non loyal client? Should the client receive a discount if it is known that (s)he is likely buy on credit and then default? etc) if data information from both sides are shared. We are currently ironing out some data limitations that we have obtained from the company and if we can work around those using econometric modeling. We have received

very rich data but are currently missing data on rejected purchases from the financial department. This leads to an abandoned shopping card. Without that data we have a problem of non-random missing data. There are statistical techniques for this to work around but we are trying to get the observed data instead. The company traditionally does not store this data, but there may be ways around it.

Project 2: This project is in collaboration with LinkedIn, who is the largest professional social network in the world. For instance, LinkedIn makes more than 50% of its revenue from providing hiring solutions to job seekers (source: LinkedIn financial report 2012). One of the key challenges for LinkedIn is to identify job seekers, as most job seekers will not publicly announce they are seeking for a job. However, job seeking behavior can be indirectly observed through how job seekers use LinkedIn.

In this research we are using longitudinal individual-level social network and site activity data from a large sample of LinkedIn members to identify job seekers. We built a hidden Markov model (HMM) in which the different states correspond to different levels of job seeking, where each state is characterized by a multivariate set of behaviors in the social network site. The model allows us to identify the members' activities on LinkedIn that most likely reflect their job seeking status. We use the model to predict the likelihood of each member's job seeking status at any point in time.

Our approach allows for early detection of job seekers based on their various activities on the social network site. For instance, we demonstrate that our predictions can be used for targeting job seekers with InMails leading to a 50.3% increase in revenue for LinkedIn over current targeting practices for the InMail recruiting product.

We are currently in the process of writing up the paper. The target journal is Management Science. We expect to submit the paper in Fall 2014. The past year the project has been presented at several conferences, among which:

Ebbes, P., O. Netzer, "Using hidden Markov models to identify job seekers from social network data", Joint Statistical Meeting (special session), August 2013

Ebbes, P., O. Netzer, "Using hidden Markov models to identify job seekers from social network data", ART Forum, June 2013

Ebbes, P., O. Netzer, "Using hidden Markov models to identify job seekers from social network data", Theory + Practice in Marketing, London Business School, May 2013

Ebbes, P., O. Netzer, "Using hidden Markov models to identify job seekers from social network data", HEC-ESSEC-INSEAD seminar, February 2013

Ebbes, P., O. Netzer, "Using hidden Markov models to identify job seekers from social network data", Marketing Dynamics Conference, Tilburg, August 2012

Ebbes, P., O. Netzer, "Using hidden Markov models to identify job seekers from social network data", Marketing Science Conference, Boston, June 2012.

Project title: Observational Learning in Internet Markets
Principal Investigator: Manasa Patnam (CREST-ENSAE)
Co-authors: Christian Helmers (Universidad Carlos III de Madrid), Pramila Krishnan (University of Cambridge)

In our research we ask whether online retail platforms successfully influence consumers' search process by providing specific product recommendations. Specifically, we examine the effect of a product referral on the referred product's sales in an online market. We use daily transaction-level data from an online fashion retailer which offers non-personalised product referrals for every product in their catalogue. To isolate the effect of product referrals, we rely on the timing of new product arrivals and the fact that new products are highly salient around the time that they are introduced. We examine the effect of being recommended by a new product on subsequent sales of the recommended product. We also exploit regional differences in the set of recommended products to build a counterfactual. We find a 5% increase in sales of existing products after they are recommended by a new product. Additionally, new products that are recommended by other new products see a 20% increase in their sales relative to new products that are not recommended by other new products. Our counterfactual analysis that relies on regional variation in product referrals confirms our results. Products that are recommended by a new product exclusively in the European region see no change in their American sales.

Project title: Bayesian Nonparametric
Principal Investigator: Judith Rousseau (CREST-ENSAE)
Coauthors: Sophie Donnet (INRA, on leave in Mexico during the period), Vincent Rivoirard (Dauphine) and Catia Scricciolo (Bocconi University)

During the year 2013-2014, in the context of ECODEC, I have worked on

- Posterior concentration rates for empirical Bayes, where we prove a general Theorem and apply it to non trivial nonparametric mixture models, this has lead to a preprint submitted at the Annals of Statistics. This is a joint work with Sophie Donnet (INRA, on leave in Mexico during the period), Vincent Rivoirard (Dauphine) and Catia Scricciolo (Bocconi University)
- Posterior concentration rates for Aalen counting processes , which has lead to a preprint submitted to Bayesian Analysis. This is a joint work with Sophie Donnet (INRA, on leave in Mexico during the period), Vincent Rivoirard (Dauphine) and Catia Scricciolo (Bocconi University)
- Asymptotic behaviour of the Bayes factors for goodness of fit tests in non iid examples , submitted to JSPI, this is a joint work with Taeryon Choi, Korea National university.
- Bernstein von Mises Theorem for semi parametric models, under revision in the

Annals of Statistics . This is a joint work with Ismael Castillo.

- I have also worked with C.Holmes on the construction of a sequential Bayesian testing procedure for the 2 sample problem. The manuscript is under preparation, but a proceeding has been written and published on a special issue of Metron for the SIS - conference .

- Conferences financed by Ecodec

- SiS Cagliari (National Italian conference, invited) June 2014 : Bayesian nonparametric tests
- ISBA world meeting, Cancun -invited conference : Bayesian nonparametric tests.
- O'Bayes for Zoe van Havre , PhD student. Duke University USA dec 2013
- ISBA world meeting for JB Salomond , PhD Student.

- Visits financed by Ecodec

- Visit of Johannes Schmidt-Hieber [Leiden Univ, Netherlands], one week, april 2014
- Visit Natalia Bochkina [Edinburgh univ, UK] , one week 2013
- Visit Kerrie Mengersen [QUT , Australia] 3 days

- Preprints

- M. Hoffman, J. Rousseau and Johannes Schmidt-Hieber (2013): On adaptive posterior concentration rates. *Revision in the Annals of statistics*
- I. Castillo and J. Rousseau (2013). A general Bernstein von Mises Theorem in semi-parametric models. *Revision in the Annals of statistics*
- S.Donnet, V.Rivoirard, J.Rousseau and C.Scricciolo (2014) Posterior concentration rates for Aalen counting processes. Submitted
- S.Donnet, V.Rivoirard, J.Rousseauand C.Scricciolo (2014)Posterior concentration rates for empirical Bayes procedures, with applications to Dirichlet Process mixtures. Submitted

Project title: The Impact of the Kobe Earthquake on Manufacturing : Analysis using Plant Level Data and Damage Maps
Principal investigator: Eric Strobl (Ecole Polytechnique)
Co-authors: Toshihiro Okubo (Keio University)
John Mutter (Columbia University)
Robert Elliott (University of Birmingham)

Project 1: The goal of the project is to assess the impact of the Kobe earthquake on the births, survival, and productivity of manufacturing plants. To this end the damage maps have now been geoprocessed, and the manufacturing plants geo-localized. Some preliminary analysis regarding the impact on plant behaviour has now been undertaken and a first draft of a paper examining this aspect has now been submitted to the Economic Journal. Further work includes purchasing and cleaning data on the financial aspects of the firm and linking these to the

damage maps. Additionally, we are in the process of purchasing high resolution monthly nightlight image data to verify our damage maps, as well, as examine how general economic activity and electricity use not captured in our manufacturing data had been affected by the earthquake.

The first paper from the project is entitled « Natural Disasters and the Death, Life, and Birth of Plants:The Case of the Kobe Earthquake »

Abstract : We examine the impact of the 1995 Kobe earthquake on the survival of manufacturing plants, their post-earthquake economic performance, and the birth of new ones. Using geo-coded plant location and unique building-level surveys we are able to identify for the first time the actual damage to the building where each plant was located in at the time of the earthquake. Including plant and building-characteristics as well as district-level variables to control for spatial dependencies, our results show that damaged plants were more likely to fail than undamaged plants and that this effect persisted up to 8 years for some. Plant survival was also adversely affected by local building and infrastructure damage. Further analysis shows evidence of falling total employment and value added associated with earthquake damage for survivors. However, we find some evidence of creative destruction with the average surviving plant experiencing a time limited increase in productivity following the earthquake. On average earthquake damage tended to deter plant births, although severe damage in an area appears to have acted as a stimulus.

Use of funds: The funds have thus far only been used to present the paper at two conferences (European Trade Study Group Conference and Verein fuer Sozialpolitik Conference). The remainder of the funds will be used to purchase data, visit Columbia University to work with John Mutter and University of Birmingham on the data to be purchased and that already in our hands.

Project title: Estimation and prediction in models described by high-dimensional matrices
Principal investigator: A. Tsybakov (CREST-ENSAE)
Co-author: Olga Klopp (CREST-ENSAE)

My research during this period was mainly devoted to errors-in-variables model in high-dimensional setting when the number of covariates can be much larger than the sample size (joint work with A.Belloni and M.Rosenbaum). We show that under suitable sparsity assumptions, the Compensated MU selector is almost optimal in a minimax sense and can be efficiently computed via a linear programming routine. Furthermore, we provide an estimator which attains the minimax efficiency bound. This estimator is determined by a second order cone programming minimization problem that can be solved numerically in polynomial time.

Publications and preprints:

1 Belloni, A., Rosenbaum, M., Tsybakov, A.B. (2014) Linear and convex programming estimators in high-dimensional errors-in-variables models. arxiv1408.0241

2 Tsybakov, A.B. (2014) Aggregation and minimax optimality in high-dimensional estimation. In: Proceedings of the International Congress of Mathematicians, Seoul, August 2014, to appear.

3 Rakhlin, A., Sridharan, K., Tsybakov, A.B. (2014) Empirical entropy, minimax regret and minimax risk. Accepted to Bernoulli.

4 Kerkyacharian, G., Tsybakov, A.B., Temlyakov, V., Picard, D., Koltchinskii, V. (2014) Optimal exponential bounds on the accuracy of classification. Constructive Approximation, v.39, 421-444.

5 Dalalyan, A., Ingster, Y., Tsybakov, A.B. (2014) Statistical inference in compound functional models. Probability Theory and Related Fields, v.158, n.3-4, 513-532.

6 Gautier, E., Tsybakov, A.B. (2013) Pivotal estimation in high-dimensional regression via linear programming. In: Empirical Inference – Festschrift in Honor of Vladimir N. Vapnik, B.Scholkopf, Z. Luo, V. Vovk eds., 195 -204. Springer, New York e.a.

7 Rosenbaum, M., Tsybakov, A.B. (2013) Improved matrix uncertainty selector. In: From Probability to Statistics and Back: High-Dimensional Models and Processes – A Festschrift in Honor of Jon

A. Wellner, M.Banerjee et al. eds. IMS Collections, v.9, 276-290. Institute of Mathematical Statistics.

Invited talks at the international conferences and invited lecture series:

2013 -Workshop ”Dependent functional data”(Göttingen, 24.01.2013-26.01.2013)

2013 -International Workshop on Statistical Learning (Moscow, 26.06.2013-28.06.2013)

2013 -Invited lectures Saint-Flour Course ”Aggregation and High-dimensional Statistics” (06.07.201320.07.2013), 12 hours.

2013 -Joint Statistical Meetings (Montreal, Canada, 03.08.2013 -09.08.2013), 30 min invited session talk.

2013 -Conference on Structural Inference in Statistics (Potsdam, 17.09.2013-19.09.2013).

2014 -SAMSI Workshop ”Statistical Inference in Sparse High-dimensional Models: theoretical and computational challenges” (Durham, USA, 24.02.2014-26.02.2014)

2014 -Journées de Statistique (Rennes, 02.06.14 -06.06.14, 2014)

2014 -2nd Conference of the International Society for Nonparametric Statistics (Cadiz, 12.06.14 16.06.14)

2014 -Conference ”Probability Theory and Statistics in High and Infinite Dimensions” (Cambridge, UK, 23.06.14 -25.06.14)

2014 -Invited short course ”Aggregation and High-dimensional Statistics”, Universit¨at Mannheim (10.07.2014-24.07.2014), 6 hours.

2014 -Invited Lecture at the International Congress of Mathematicians, Seoul (14.08.14 -21.08.14)

Honors:

Gay-Lussac – Humboldt Prize (2013)

5 Invited professors

- Kirill Evdokimov (Princeton, 2013) : collaboration with E. Gautier, discussion with researchers of the team and students, presentation at a seminar
- Alexandre Nazin (Academie des Sciences Russe, 2013) : collaboration with A. Tsybakov, discussion with researchers of the team and students, presentation at a seminar
- Bin Yu (Berkeley, 2013) : collaboration with A. Tsybakov, , discussion with researchers of the team and students, presentation at a seminar
- Felix Abramovich (Tel Aviv, 2014) : collaboration with A. Tsybakov, discussion with researchers of the team and students, PhD lecture (OFPR cours): “Model selection in high-dimensional regression and related issues”
- Andrew Gelman (Columbia, 2014) : collaboration with N. Chopin and J. Rousseau, discussion with researchers of the team and students, lecture (master level): “Bayesian data analysis”

6 PhD and postdoc grants

- Pierre Bellec (2013-2016) : PhD with A. Tsybakov, « Agrégation d'estimateurs de densité: optimalité, parcimonie en grande dimension et universalité »
- Quentin Paris (2013-2014) : post-doctorate with E. Gautier : « Modèles de choix multiples avec coefficients aléatoires, relaxations convexes et demande pour produits différenciés »
- Boton Szabo (6 months in 2014) : post-doctorate with J. Rousseau: “Estimation Bayésienne non-paramétrique - propriétées fréquentistes de méthodes dites bayésiennes empiriques »
- Benjamin Poignard (01/10/2014 – 30/09/2017) : PhD with Jean-David Fermanian: “Modèles dynamiques de matrices de corrélations à base de Vines”
- Jean-David Sigaux (01/09/2014 – 31/08/2015): PhD “Three Essays on Bond Lending”
- Martin Thorsten (01/09/2014 – 31/08/2015): PhD “Essays in Financial Economics”

7 Conferences

- "Nonparametric and high-dimensional statistics », Centre International de Rencontres Mathématiques de Marseille, December 2012, (co-funding and co-organized by some members of Axe 5).

- "Stats in Paris: a school and conference on statistics and econometrics of networks", ENSAE, 18-22 November 2013 (co-funding and co-organized by some members of Axe 5).

Stats in Paris School and Conference was co-financed by Labex Ecodec, Labex MME-DII, Centre for Microdata Methods and Practice (London) and Ensae ParisTech.

Three Research courses were proposed
L. Ménard (Paris 10): Probabilistic foundations of graphs and networks
S. Goyal (Cambridge): Economics of networks
E. Kolaczyk (Boston University): Statistical analysis of network data

The conference took place during two days and gathered 13 speakers and 150 participants, 50 % being from foreign countries.
10 mobility grants have been given to young researchers.
27 posters have been selected and presented in flash poster sessions.

- "Building Social Media Intelligence" (organized by Peter Ebbes and Francesca Sotgiu): series of 3 workshops. The first workshop of this series has been held on May 18 and 19 2014 by David Schweidel (Emory University), the second was held on 28 and 29 August 2014 by Pete Fader (University of Pennsylvania) and Bruce Hardie (London Business School) and the third took place on 18 and 19 September 2014 with Eva Ascarza and Oded Netzer (Columbia University) on Hidden Markov Models in Marketing.

- "Data Lead 2014", Berkeley, 30 September - 2 October, 2014 (co-funding and co-organized by some members of Axe 5).

This conference was organized and co-financed by Labex Ecodec, The Haas School of Business (University of Berkeley) and Ensae ParisTech.

Workshop by David Schweidl (Emory University) on Building Social Media Intelligence (May 18 and 19 2014)

Principal organizer: *Peter Ebbes (HEC Paris)*

Summary of completed work:

The workshop went very well. We had 27 participants in the room in the end while we targeted 25, so there was a little more demand than foreseen. The workshop went over very well from what I heard from the participants, despite the big difference in audience background (ranging from faculty, PhD to MBA and GE). David (the teacher) did an excellent job discussing something that was useful to all of us in the audience, and everyone learned many new things. David himself also really liked the format, where one day was fully devoted to a workshop and the 2nd day fully devoted to research.

The workshop was targeted to quantitative students (Ph.D., MBA, Grand Ecole, Master) with solid analytical skills and strong interest in social media and digital analytics, as well as all interested faculty.

Workshop abstract:

In the world of Facebook, Twitter and Yelp, water-cooler conversations with co-workers and backyard small talk with neighbors have moved from the physical world to the digital arena. In this new landscape, organizations ranging from Fortune 500 companies to government agencies to political campaigns continuously monitor online opinions in an effort to guide their actions. Are consumers satisfied with our product? How are our policies perceived? Do voters agree with our platform?

But measuring online opinion is more complex than just reading a few posted reviews. Social media is replete with its share of noise and chatter that can contaminate monitoring efforts. But by knowing what shapes online opinions, we can better uncover the valuable insights hidden in the social media chatter – insights that can inform our organization's strategy.

In this workshop, we discuss how organizations can move beyond the current practice of social media monitoring to develop social media intelligence that can drive marketing decisions.

Workshop by Pete Fader (University of Pennsylvania) and Bruce Hardie (London Business School) on probability models for Customer Lifetime Value (CLV), (August 28 and 29 2014)

Principal organizer: *Peter Ebbes (HEC Paris)*

Summary of completed work:

The workshop went very well. We had 36 participants in the room in the end while we targeted 25, so there was a more demand than foreseen. The workshop went over very well from what I heard from the participants, despite the big difference in audience background (ranging from faculty, PhD to MBA and GE). Pete and Bruce did an excellent job discussing something that was useful to all of us in the audience, and everyone learned many new things. Both Pete and Bruce were impressed by the audience and their engagement, and they liked the format of the two days, where one day was fully devoted to a workshop and the 2nd day fully devoted to research.

The workshop was targeted to quantitative students (Ph.D., MBA, Grand Ecole, Master) with solid analytical skills and strong interest in social media and digital analytics, as well as all interested faculty.

Workshop abstract:

Over the course of many decades, applied statisticians have developed a number of models that have proven to be highly effective in their ability to explain and predict behaviour in many areas of business (and, more generally, the social sciences). These models use some basic "building blocks" from probability theory to offer behaviourally plausible perspectives on different types of timing, counting, and choice behaviours, and have been part of the marketing science literature from the very beginning. Furthermore, they are the foundations on which many of today's "leading-edge" marketing models have been built.

The objective of this workshop is to familiarise participants with the core probability models used by marketing scientists and to show how they can be applied to practical marketing problems. The context in which we will introduce these models is that of computing customer lifetime value (CLV). At the heart of any calculation of CLV is the generation of multi-period forecasts of buyer behaviour. We will explore how to derive, implement (in Excel), and validate simple probability models that can generate such forecasts, and thereby compute CLV in different business settings.

Workshop by Eva Ascarza and Oded Netzer (Columbia University) on Hidden Markov Models in Marketing, September 18 and 19 2014

Principal organizer: *Peter Ebbes (HEC Paris)*

Summary of completed work:

The third workshop in the series was the best attended. We had 44 participants in the room in the end while we targeted 25, so there was much more demand than foreseen. The workshop went over very well from what I heard from the participants, despite the big difference in audience background (ranging from faculty, PhD to MBA and GE). Oded and Eva did an excellent job discussing something that was useful to all of us in the audience, and everyone learned many new things. Both Oded and Eva were impressed by the audience and their engagement, and they liked the format of the two days, where one day was fully devoted to a workshop and the other day fully devoted to research.

The workshop was targeted to quantitative students (Ph.D., MBA, Grand Ecole, Master) with solid analytical skills and strong interest in social media and digital analytics, as well as all interested faculty.

List of affiliation and program of participants:

As with the other two workshops, what is important to note is the broad audience (i.e. background) but also that we reach a large set of participants across Europe. This event really helps putting HEC Paris on the (European) map in the area of big data.

Workshop abstract:

Customers or the business environment often transition over time from one mode of behavior to another. For example, a customer may transition from a more positive relationship with the firm to a less positive one due to exposure to an attractive competitor. However, managers often do not observe the customer's buying behavior state, neither the transitions from one state to another. Rather, they need to infer such changes from the customer's observed behaviors (e.g., purchases, transactions, searches). Hidden Markov models (HMMs) are valuable tools for such situations. More generally, HMMs are useful for situations in which the marketer or researcher is interested in identifying a (dynamic) latent state of the world from a series of (possibly noisy) observations. Recently, these models have been widely applied to marketing problems. In marketing, firms are often interested in understanding their customers' latent buying behavior state and assessing how the firm can use its marketing levers to move customers to a more favorable state to the firm.

In this workshop we will discuss what are hidden Markov models, what are their advantages and perils and how should one go about applying such models to marketing data. We will demonstrate these topics using examples of applications of these models in marketing and related fields. The workshop will also include hands on estimation of hidden Markov models using R. No pre-existing knowledge of R will be assumed, though participants are encouraged to install the R statistical software (freely available from http://www.r-project.org/) on the laptops prior to the workshop.

Next conferences to be organized

- "Journées ENSAE/ENSAI" (organized by A. Simoni and N. Klutchnikoff). Period: January 8-9, 2015.

- "Conference on Learning Theory" (organized by Gilles Stoltz and : Vianney Perchet). Period: June 2015. Plenary speakers will be Cédric Villani, Lyon (Médaille Fields 2010), Dan Spielman, Yale (Prix Gödel 2008), Tim Roughgarden, Stanford (Prix Gödel 2012).

- "Sequential Monte Carlo conference" (organized by N. Chopin). Period: 26 to 28 August 2015, CREST-ENSAE.

- "Bayesian Nonparametrics meeting". Period: June 2016